Bangla RST Discourse Treebank: Annotation Guidelines

Manfred Stede
University of Potsdam
stede@uni-potsdam.de

Debopam Das
University of Potsdam
debdas@uni-potsdam.de

Version 2.0 (January 2018)

Notes:

- Parts of this document are a translation of the Chapter 'Rhetorische Struktur' from the collection of annotation guidelines in Stede (2016).

- These guidelines explain the procedure of RST analysis only in an abstract manner (for "on paper" analysis). Hints on using a specific annotation tool will be provided separately.

- All linguistic examples for which no source is provided are fictitious. Our English corpus examples are from the RST Discourse Treebank (RST-DT) (Carlson et al., 2003) and translations from the Potsdam Commentary Corpus (PCC) (Stede and Neumann, 2014). The Bangla examples are from the Bangla RST Discourse Treebank or Bangla RST-DT (Das and Stede, to appear), which includes newspaper texts from a Bangla daily named *Anandabazar Patrika*.

- Thanks are due to Gisela Redeker for discussion of these guidelines and suggestions for improvement.

1 Background

Rhetorical Structure Theory (Mann and Thompson, 1988, Taboada and Mann, 2006)[1] postulates that a coherent text can be characterized by means of a tree structure whose leaves are the "elementary discourse units" (henceforth: EDUs), and whose internal nodes are labelled with a *coherence relation*. The theory proposes a list of about 25 such relations, which may be said to hold between two EDUs or (recursively) between larger spans of text. With most of the relations, the two units are not of equal "weight". Instead, one of them is said to play the central role of this local constellation, while the other one has merely a supportive function. The former unit is called the *nucleus* of this (instance of a) coherence relation; the latter is its *satellite*.

For illustration, Figure 1 shows the RST analysis of a short piece from a text, which has been produced using the *RSTTool*[2] software. Horizontal lines denote

---

[1] These two papers are recommended as background reading for annotators.

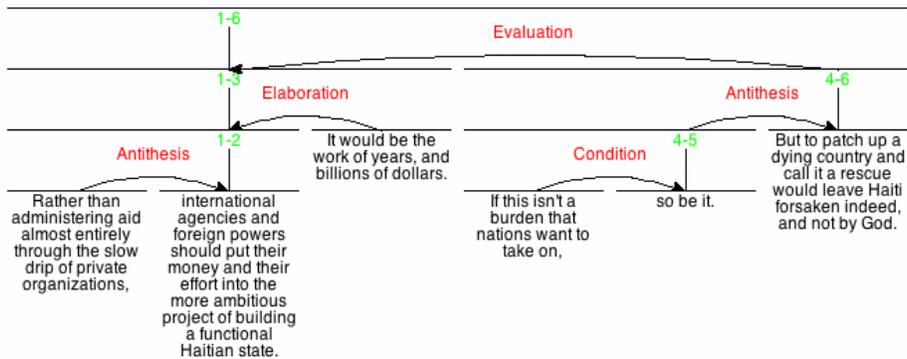[2] http://www.wagsoft.com/rsttool (accessed: Sep 30, 2015)

Figure 1: Example of an RST analysis of a text excerpt (Source: George Packer: Suffering. *The New Yorker 42, 2010*)

text units: either EDUs, corresponding to the leaves of the tree, or sequences of EDUs forming larger spans. Curved arrows connect the satellite of a relation to its nucleus, which furthermore is marked by a vertical line that connects it to its embedding larger segment. The names of the relations are given in red.

What is the status of such an anaysis? We define it as the *reconstruction of the author's plan from the perspective of the reader*. Under this view, RST is a tool for the explication of the reader's text understanding: The tree structure demonstrates to a good extent the intention that the author supposedly had when producing the text. Notice that this view implies that an RST analysis by definition includes a certain amount of subjectivity – different readers might very well produce somewhat different analyses, which can nonetheless be equally legitimate or "valid".

2   Goal of the annotation

The rhetorical structure of a text describes how adjacent spans of units are connected to each other. This may be a relatively loose thematic "continuation" (e.g., using the coherence relation ELABORATION) or a very specific kind of connection, as for example in a *if…then* structure corresponding to the relation CONDITION. The set of coherence relations will be described in detail below in Section 4. Overall, an RST analysis consists of four tasks:

- Break the text into elementary discourse units; this step will be explained in Section 3.

- Decide on the *hierarchical structure* of the text: Which adjacent units are to be connected to each other in what order – what is the resulting tree structure that covers the complete text?

- When joining two adjacent units into a larger one, decide on the coherence relation to be applied.

- Decide on the *nucleus/satellite* status of the linked units. For most relations, this follows automatically from their definition. Some relations come in two variants, allowing to select the nucleus freely. In general, the overall text structure that is being created can impose constraints on the nuclearity assignment – which thus may put some restrictions on what relations can or should be used.

In practice, the last three tasks are not performed separately but are closely tied to one another. At the end of the annotation process, the complete text has to be covered by the tree structure, without any gaps (EDUs that are not participating in the analysis). At any point, adjacent units are being related to one another such that no crossing edges originate in the tree. (In other words, the tree is "projective".)

Another important property of the tree is that no node has more than one parent node, which means that any unit of the text can play only one role in the rhetorical structure. It thus cannot function as a satellite in two distinct relations, being linked to different nuclei. On the other hand, it is possible that a single nucleus has multiple satellites, as it is illustrated with the example in Figure 2. The final sentence amounts to the central statement of the text and therefore constitutes the *central nucleus* of the text: If you start at the top (root) node of the tree and move to that leaf, along the path you encounter only nucleus links.

The motivation for that sentence results directly from sentence 2, which is—according to this analysis—the second-most important one in the text. We can check this with a paraphrase text, according to which a sequence of the "nuclear" EDUs of the text, connected in a way that reflects the coherence relations, should result in a summary of the text: "Since you should be careful when camping in the wilderness, do not store any meat in your tent."

In the analysis, the MOTIVATION relation holds between sentences 1–4 (satellite) and 5 (nucleus). The structure now postulates that it automatically also
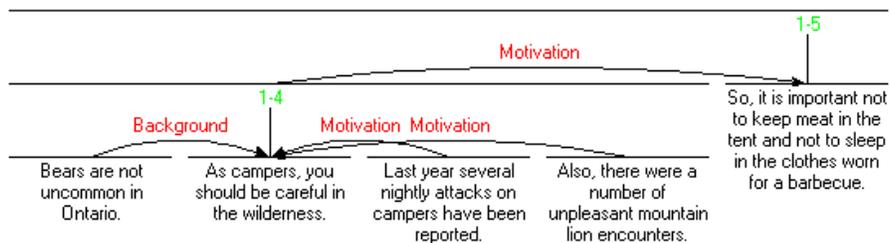

Figure 2: Another example of an RST analysis

3

holds between the nucleus of 1–4 (i.e., 2) and 5. This property of RST trees has been formulated as the *strong nuclearity principle* by Marcu (2000). It is particularly relevant when analyzing argumentative texts.

For sentences 1–4, the analysis sees no further hierarchical structure; instead, it connects 1, 3 and 4 individually to 2. One might argue that the first three sentences deal with the topic of bears, and therefore 1 should be the BACKGROUND of 2–3, but sentence 4 also connects to 2 (in a parallel way), so that building a unit for 1–3 would be counterintuitive for the structure of the text.

As we have seen, the different spans of the text receive different "weights" by assigning nuclearity status to EDUs and larger units. In the following, we call the central units of the text "strongly nuclear": The path connecting them to the root of the tree includes no or only very few satellite links.

The next section explains the preparatory step of segmentation; afterwards, Section 4 introduces the inventory of relations and illustrates them with examples. Section 5 explains the concrete procedure for performing a text analysis. Finally, Section 6 discusses the various annotation decisions and their interactions by means of a complete analysis of a text from the RST Discourse Treebank (Carlson et al., 2003).

3   Segmentation

The segmentation guidelines followed in the Bangla RST-DT are based on the guidelines used for German texts in the Potsdam Commentary Corpus or PCC (Stede, 2016) and for English texts in SLSeg (syntactic and lexically based discourse segmenter) (Tofiloski et al., 2009). These guidelines were also used in Das et al. (2017) for segmenting German and English texts, respectively.

Both PCC and SLSeg guidelines closely adhere to the original definition of spans in RST, which considers that only adjunct, and not complement clauses, form legitimate EDUs. This principle specifies that every EDU must contain a verb, either finite or non-finite. Broadly, we consider coordinated clauses (but not coordinated verb phrases), adjunct clauses and non-restrictive relative clauses (separated from the main clause by a comma in English) to establish legitimate EDUs. In all cases, this strategy is, however, complemented by the annotator's decision on whether a discourse relation could hold between the resulting segments.

As we primarily follow formal criteria for determining the status of EDUs, we closely examine how clausal structures are realized in Bangla. For this purpose, we look into the existing literature on the Bangla grammar, and consult some notable works such as Chatterji (1988), Chakraborty (1992), Chaki (1996) and Sarkar (2006), which altogether provide a comprehensive account of clausal constructions in Bangla.

In the following subsection, we enumerate specific guidelines used for seg-

menting texts in the Bangla RST-DT. Most of the examples (accompanying specific guidelines) are taken from the Bangla RST Discourse Treebank (Das and Stede, to appear). The example sources (file numbers) are mentioned at the end of each example. If there is no file number, then the example is an invented one. The text within a pair of square brackets denotes an EDU. The text in the Bangla examples is written in the Roman script (ITRANS style).

3.1 Segmentation guidelines for Bangla

3.1.1 Zero-copula Constructions

Bangla allows frequent uses of zero-copula constructions, in which the main copular verb (corresponding to the verb 'be' or 'have' in English) remains overtly absent, but in effect, is implied. Although in RST segmentation, a legitimate EDU is required to contain a verb, we decide to consider zero-copula constructions as clauses (headed by an implicit, but implied verb) and hence as EDUs, unless they act as complement clauses of other verbs.

(1) [sAjid o   pArbhin svAmI-strI.]
    Sajid and Parvin   husband-wife

    Sajid and Parvin are husband and wife. [kolkata-05]

We would also like to point out that Bangla is a pro-drop language, in which subject pronouns are omitted from clauses on many occasions. In our annotation, we consider such constructions (clauses only with verbal predicates, and not the overt subjects) as EDUs.

(2) [er   par Ar   pratiyogitAmulak Asare   nAmben
    this.Gen after anymore competitive   tournament will participate
    nA.]
    not

    (He) will not participate in competitive tournaments anymore after this. [sports-03]

3.1.2 Clausal Subjects

Clausal subjects are not considered to be EDUs. In Bangla, clausal subjects are often manifested by verbal nouns.

(3) [upayukta sarkAri   bandobasta thAkA jaruri.]
    proper   governmental provision   be   necessary

    The proper governmental provision is necessary. [editorial-column-08]

Sometimes, a complete clause (with a finite verb) can also be used as the subject of a sentence.

(4) [se   bandobasta ekebArei   nei, emanTA sambhabata balA yAbe
such provision   at all.Emph not that   probably   say   can
nA.]
not

That there is no such provision at all cannot be said. [editorial-column-08]

3.1.3 Clausal Complements

Clausal complements include clausal objects of verbs, expressed as verbal nouns (Example 5) or infinitival clauses (Example 6), and they are not considered to form EDUs.

(5) [bahu mAnuSh dAktArer chembAre   jAoYAr cheYe jyotiShIr
many people   doctors'   to chamber go.Gen   than   astrologers'
chembAre   jAoYA beshi paChanda karen.]
to chamber go   much prefer   do

Many people prefer to go to astrologers' chambers than doctors' chambers. [letters-to-the-editor-06]

(6) [jiesTi kiChuTA hAsi phoTAte chaleChe bAik   bhaktoder
GST   a little   smile to bring go.Prog   motorcycle fans'
mukheo.]
face.Emph

GST is also going to bring a little smile on the faces motorcycle fans. [business-06]

3.1.4 Attribution Clauses

Attribution clauses are often represented by reported speeches, both directly (by direct quotes) or indirectly. We believe that attribution is a syntactic phenomenon, rather than a discourse one. Since attribution clauses act as the complements (more like noun clause complements) of the main reporting verbs within a matrix clause, they are not assigned the status of EDUs.

(7) [praphesar AYAn hoYAT boleChen, "ekhonai   Ata.mkita haYe
professor   Ian   Howat said   now.Emph panicked   be
parar   konao   kAron nei"]
get.Gen any.Emph reason not

Professor Ian Howat said, "There is no reason to get panicked by now." [science-04]

(8) [goYendAprodhAn   Aro jAnAn, dhritader   jiGYAsAbAd
the chief of detectives more informed arrested ones' interrogation
karA hochChe.]
do   be.Prog

The chief of detectives also informed that the arrested ones are being interrogated. [kolkata-05]

Another way attribution clauses can manifest themselves is through cognitive predicates (containing verbs expressing feelings, thoughts or opinions, such as *think*, *know*, *estimate* or *wonder* in English). Just as in the case of reported speeches and for the similar reason, cognitive predicates are not treated as EDUs in our annotation.

(9) [hAmlAr   prAthamik laxya tA.Nr bA.Dii Chilo bale sandeha
     of the attack primary   target his   house was   that suspicion
     karChen tadantakArIrA.]
     do.Prog investigators

     The investigators are suspecting that the primary target of the attack was his house. [international-01]

We generally do not segment between the reporting clause and the reported clause, or between the reported clauses. However, for longer speeches consisting of multiple sentences, we have observed that if we strictly follow this principle, we might end up losing significant information at the discourse level. Thus, we have decided to add an exception: If a reported speech (or quote) spans over more than one sentence, then each sentence will be segmented as EDUs (marked by square brackets in Example 10).

(10) "[bhAloi   haYeChe daurTA.]   [Ami saThik
      good.Emph has been the (sprint) race I   right
      pathei   yAchCHi.] [tabe   ekhanao anek kAj
      in-the-direction.Emph moving   However still   many things
      bAki.",   baleChen bolT.]
      remaining

      "The (sprint) race has been good. I am moving in the just right direction. However, there are still many things to do.", said Bolt. [sport-03]

3.1.5   Relative Clauses

Relative clauses in Bangla are represented by correlative pronouns, sometimes in reduplicated forms (e.g., *ye / se, yini / tini, yata / tata, yArA yArA / tArA, yekhAne yekhAne / sekhAne sekhAne*). We exclude restrictive relative clauses from our consideration of EDUs. This is because we believe that coherence relations only hold between propositions, rather than between entities or between an entity and a proposition.

(11) [jini lulAr sAjA   ghoShanA karlen, tinio   rAjnItite Aste
      who Lula's sentence announced   he.Emph in politics to come
      AgrahI.]
      interested

      He who announced the sentence of Lula is also interested to join politics. [international-05]

However, non-restrictive clauses are considered to be EDUs in our annotation, because they enter into a coherence relation with the propositions expressed by their host clauses.

(12)   [sirAj je   mirjApharer upar bharsA koreChilen,] [seTA pore tAr
       Siraj that Mirzafar's   on   relied   that later his
       pataner   kAran haYe d.NA.DAY]
       downfall's reason be   stood

       Siraj relied on Mirzafar, which later became the reason of his downfall.

Unlike in English, both restrictive and non-restrictive relative clauses in Bangla can be separated from the main clause by a comma.

3.1.6   Clauses with Correlative Discourse Connectives

In addition to correlative pronouns (for relative clauses), Bangla also contains correlative discourse connectives (sometimes in reduplicated forms) which are used to connect two clauses. Examples of correlative connectives include *yehetu / sehetu, yeman (yeman) / teman (teman), yadi / tabe,* etc. Clauses with such connectives are considered to be EDUs in our annotation.

(13)   [yadi tumi Aso,] [tabe Ami yAbo.]
       if   you come then I   will go

       If you come, then I will go.

3.1.7   Nominal Modifiers

Nominal modifiers represented by verbal nouns are not considered as EDUs.

(14)   [ulTo   dik theke AsA bAsTike dhAkkA mAre oi   gA.DiTi.]
       opposite side from come the bus hit   that car

       The car hit the bus coming from the opposite side. [international-01]

3.1.8   Participial Clauses

Participial clauses (with a past active participle), are considered to constitute legitimate EDUs.

(15)   [dvitIYa TesTe   phire   ese] [sirij 1-1 karlen phAph
       second   in the test coming back   series 1-1 did   Faf
       duplesi.]
       du Plessis

       Coming back in the second test, Faf du Plessis made the series 1-1.
       [sport-08]

3.1.9 Verbal Nouns with a Postposition

Verbal nouns with a postposition are considered to form EDUs.

(16) [eman sambhAbanAder chine   neoAr janya] [upayukta
     such potentials   recognize.Gen   proper governmental
     sarkAri   bandobasta thAkA   jaruri]
     provision be   necessary

     The proper governmental provision is necessary for recognizing such po-
     tentials. [editorial-column-08]

3.1.10 Infinitival Clauses

Infinitival clauses which are not complements of verbs are considered as EDUs.

(17) [nyAnoke bhabiShyate rAstAy chAlAte] [dubaCharer madhyei
     Nano   in the future on road run   of two years within.Emph
     chAi natun lagni.]
     want new   investments

     New investments are required within the next two years in order to run
     Nano on road in the future. [business-05]

3.1.11 Conditional Clauses

Conditional clausal constructions in Bangla act as adjunct clauses, and hence
they are considered to form EDUs.

(18) [jiesTir parimAn kam hale] [sexetre dAm kambe   gA.Dir]
     GST's   amount small be.if then   price will go down cars'

     If the amount of GST is small, then the price of cars will go down.
     [business-06]

3.1.12 Coordinated Constructions

As in many other RST annotation studies, we also consider as EDUs only coordi-
nated clauses (linked by a comma or discourse connective), but not coordinated
verb phrases.

(19) [Aphsos karChilo bA.mlA,] [Aphsos karChilo   mahAnagar.]
     regret   was doing Bangla   regret   was doing the big city

     Bangla was regretting, so was the big city. [editorial-column-11]

In sum, we followed the basic ideas of RST segmentation from the PCC
and SLSeg guidelines (for adjunct/complement clauses, attribution and relative
clauses). However, at the same time, we have developed some new segmentation
strategies suitable for certain Bangla constructions (e.g., conditional clauses).
Sometimes, we used the existing PCC and SLSeg guidelines, but have adapted

them in particular ways so that they comply with the syntactic and discourse structures of Bangla (in the treatment of relative clauses, verbal noun with a postposition, etc.).

4 Relations

The definitions of the relations used in our annotation are based on those of the RST website[3], but there are some modifications and additional hints that apply specifically to genres of opinionated text, such as editorials. Similar to the arrangement in the RST website, we introduce the relations in groups.

1. Primarily *pragmatic* relations describe (especially in editorials) the argumentation of the author: What are the claims, and how are they being supported by observations or by other claims?

2. Primarily *semantic* relations are used when the author describes a (possibly complex) states of affairs in the world; this may involve for example relations of causality among events.

3. *Textual* relations work to organize the text and make its understanding easier by providing orienting information, or repetition.

4. In all three groups above the relations hold between a nucleus and a satellite; they are thus called *mononuclear*. In contrast, the fourth group contains *multinuclear* relations, those that do not mark a difference in weight, but instead connect two or more nuclei.

The first three groups are then distinguished on the grounds of considering "content", whereas the fourth arises from a structural criterion (multinuclearity) and contains semantic, pragmatic and textual relations.

As indicated earlier, occasionally there will be no single plausible analysis for a constellation of text units. The existence of the grouping of relations is in fact one reason for this situation, as an annotator may focus more on the semantic or on the pragmatic side of the content of the units and thus favour a relation from one group over that from another one. We will provide some additional explanation after having introduced the relations, in Section 4.6.

4.1 Format of the definitions

Relations are being defined by explaining (a) the role of the two units that are being adjoined and (b) the effect that the author wants to achieve by applying the relation to the units. In addition to this core information, we provide examples (either translations from the Potsdam Commentary Corpus, or fictitious ones), mention typical connectives that can be used to signal the relation

---

[3]http://www.sfu.ca/rst (last accessed March 7, 2017). In the RST literature, this set of definitions is known as the "Extended classical relations".

in English, and finally add some remarks on applying the relations to the text genre of editorials. The overall format for the definitions is as follows:

- **N:** Description of the type and/or the function of the nucleus
  Characterizing primarily the goals of the author, not the expression in the text.

- **S:** Description of the type and/or the function of the satellite
  Characterizing primarily the goals of the author, not the expression in the text.

- **N/S:** Description of the function of the nucleus/satellite combination. When there are constraints or tendencies on the linear order of N and S, they will also be given here.

- **Effect:** Description of the effect inteded by the author in using the relation, in the form of a "before–after" change.

- **Typical connectives**

- **Example:** N and S are marked. If there is a description of prior context, it is given in *italics*.

- **Remarks:** Optional.

The fields for N, S and N/S remain empty when a relation does not have any such constraints. For multinuclear relations, the fields S and N/S are omitted.

It is important to remember that the N/S distribution can often be understood only in a wider discourse context, so that the excerpts given here are merely illustrative (e.g., in the opposition between CAUSE and RESULT and between ANTITHESIS and CONTRAST).

When sample connectives are given, please note:

- We provide merely a few examples here. Usually, there are more connectives for a given relation; sometimes there are many.

- Connectives can be ambiguous. On the one hand, many words can have another reading in addition to the connective one; on the other hand, some connectives can signal more than one relation, such as *since*, which can be used to indicate a temporal or a causal relation.

- Sometimes, a connective that "usually" signals relation X can be interpreted to mark a different relation Y. Temporal sequence and causality, for instance, often co-occur, and sometimes the causality is more important than the temporal core meaning of the connective, as in this example involving *then*:

  (20)   Little Joseph received a cookie from his grandpa and then he was finally happy.

11

During an analysis, when checking whether some relation is applicable to a pair of text segments, annotation should proceed as follows:

- Is there a connective signalling the relation, or at least constraining the set of possible relations? In some cases, however, the relation signalled at the surface is not the pragmatically central one (as illustrated above). We will discuss this further in Section 4.6.

- Punctuation generally does not provide reliable evidence for specific relations, but there are some tendencies, such as a connection between semicolon and contrastive relations:

  (21) My brother loves dark chocolate; my sister for some reason prefers the lighter milk chocolate.

- When juxtaposing the two segments, does the author want to achieve the effect that is specified in the relation definition? This is a necessary condition for applying a relation.

- Are the constraints on type or function of nucleus and satellite and on their combination fulfilled? If they are specified, they also constitute necessary conditions.

- Is the connection produced by the relation between nucleus and satellite appropriate for the larger text function? This criterion is less strict (and more prone to interpretation) than the ones given above, but it can make the decision easier when more than one relation seems applicable.

Below we first provide definitions for mononuclear relations (i.e., those where one segment is more important for the author's purposes). They are divided into a set of primarily pragmatic and semantic relations, respectively, followed by some that work to organize the text. Afterwards we describe multinuclear relations, where there is no such imbalance between the segments.

4.2 Primarily pragmatic relations

These relations cover cases where the author makes understanding of a segment easier (BACKGROUND), evokes a positive opinion on a state of affairs (ANTITHESIS, CONCESSION), justifies a thesis that the author has proposed (EVIDENCE, REASON), evaluates a state of affairs from the author's perspective (EVALUATION), attempts to trigger an action on the part of the reader (Motivation) or makes it easier for the reader to perform the action (ENABLEMENT).

Throughout the definitions, 'R' stands for 'Reader' and 'W' for 'Writer'; 'N' for 'Nucleus' and 'S' for 'Satellite'. Note that the 'Effects' are in most cases taken from the definitions of the RST Website. 'N/S' is usually more elaborate in our version.

BACKGROUND

- **N/S:** Understanding S makes it easier for R to understand the content of N; without the background information in S, it would be difficult to comprehend N. In a text, S mostly but not always precedes N. A Background S at the beginng of the text often serves to introduce the topic of the text.

- **Effect:** R's ability to comprehend N is increased.

- **Typical connectives:** Rarely signalled by connectives.

- **Example:** [Until 1984, Burkina Faso was called Obervolta.]$_S$ [According to an EMNID poll, many Europeans today believe that they are two different countries.]$_N$

- **Remark:** This relation only rarely holds between two EDUs; mostly it will connect larger segments. Many editorials start with a Background S, which serves as starting point for W's subsequent assessment.

ANTITHESIS

- **N:** W regards the content of N as more important; it is the antithesis that W is identifying with.

- **S:** In comparison to N, W regards the content of S as less important. S is considered to be the thesis which the W is not identifying with.

- **N/S:** The contents of N and S are not compatible – often on the level of their evaluation. Due to the incompatibility, one cannot have equal regard for N and S. In a text, S usually precedes N.

- **Effect:** R's positive regard for N is increased.

- **Typical connectives:** *but; neg - rather; neg - instead; …*

- **Example:** *Attempt of a city to sell off real estate* [At one point they seemed to have succeeded.]$_S$ [But the buyer didn't pay.]$_N$ (maz-6193, PCC)
  [It was not the man who received negative votes]$_S$ [but his job, which turns him into a supporter for the bombing test range.]$_N$ (maz-17300, PCC)

- **Remark:** Similar relations are Concession and multinuclear Contrast; see also Section 4.6. A subgroup of Antithesis is cases where N serves to correct an assumption stated in S.

CONCESSION

- **N:** W regards the content of N as more important.

- **S:** In comparison to N, W regards the content of S as less important, but s/he does not dispute that S holds.

- **N/S:** W concedes S and implicitly confirms that S and N are usually not compatible; in the current instance, however, they are compatible, and N is being emphasized.

- **Effect:** R's positive regard for N is increased.

- **Typical connectives:** *although; but; still; despite; ..*

- **Example:** [Sanitary facilities nowadays are a standard on big campgrounds.]$_S$ [But Radewege has a hard time with upgrading to such standards.]$_N$ (maz-6488, PCC)

EVIDENCE

- **N:** A subjective statement/thesis/claim, which R might not accept or might not regard as sufficiently important or positive.

- **S:** A statement that R is likely to accept; usually an "objective" description of a fact.

- **N/S:** Understanding S makes it easier for R to accept N, or to share the particular viewpoint of W.

- **Effect:** R's belief in N is increased.

- **Typical connectives:** Causal connectives.

- **Example:** *Debate about having two subjects 'Religion' and 'LER' at school* [And now even our state government seems determined to remove this apparent equality between the two subjects.]$_N$ [Stolpe, Reiche and others do say Yes to the possible compromise offered by the Karlsruhe court, but they also decree: There cannot be any voluntary subject area LER/Religion.]$_S$ (maz-6159, PCC)

- **Remark:** Evidence often connects a larger S segment to a shorter N (the thesis).

Reason

- **N:** A subjective statement/thesis/claim, which R might not accept or might not regard as sufficiently important or positive.

- **S:** A subjective statement/thesis/claim.

- **N/S:** Understanding S makes it easier for R to accept N, or to share the particular viewpoint of W.

- **Effect:** R's belief in N is increased.

- **Typical connectives:** Causal connectives.

- **Example:** [With each new day of air raids, the military operations of the U.S. lose more and more credibility.]$_N$ [By means of comprehensive area-wide destruction you can't hit the Taliban, nor can you eliminate bin Laden.]$_S$ (maz-5701, PCC)

- **Remark:** Reason is more specific than Evidence. The different lies on whether S is being presented by W as "objective" (Evidence) or also constitutes a subjective statement itself (Reason).

Reason-N

- **N:** A subjective statement/thesis/claim.

- **S:** A subjective statement/thesis/claim, which R might not accept or might not regard as sufficiently important or positive.

- **N/S:** Understanding N makes it easier for R to accept S, or to share the particular viewpoint of W.

- **Effect:** R's belief in S is increased.

- **Typical connectives:** Causal connectives.

- **Remark:** This relation is parallel to Reason. It applies when for the text function, the reason is more important than the cause.

Justify

- **N:** A subjective statement/thesis/claim, which R might not accept or might not regard as sufficiently important or positive.

- **S:** A statement of a fundamental (e.g., political, moral) attitude of the acting person.

- **N/S:** Understanding S makes it easier for R to accept N, or to share the particular viewpoint of W.

- **Effect**: R's readiness to accept W's right to present N is increased.

- **Typical connectives:** Causal connectives.

- **Example:** [I'll never eat this dish.]$_N$ [I'm a vegetarian.]$_S$

EVALUATION-S

- **N:** Description of a state of affairs, or a subjective statement (but not from W's perspective).

- **S:** A subjective evaluation (positive/negative, desirable/undesirable) from W's perspective.

- **N/S:** S evaluates N.

- **Effect:** R recognizes that S assesses N and recognizes the value it assigns.

- **Typical connectives:** Rarely signalled by connectives.

- **Example:** [Its past seemed to shadow the big hotel like a curse.]$_S$ [For many years the Potsdam local court had tried to find a buyer, without success.]$_N$ (maz-6193, PCC)

- **Remark:** Usually the 'evaluating' segment follows the 'evaluated' one. But sometimes the order is the opposite, as in the example.

EVALUATION-N

- **N:** A subjective evaluation (positive/negative, desirable/undesirable) from W's perspective.

- **S:** Description of a state of affairs, or a subjective statement (but not from W's perspective).

- **N/S:** N evaluates S.

- **Effect:** R recognizes that N assesses S and recognizes the value it assigns.

- **Typical connectives:** Rarely signalled by connectives.

- **Remark:** This relation is parallel to Evaluation-S. Deciding between the two depends solely on judging the relative importance of the segments for the text.

- **N:** An action to be performed by R.

- **N/S:** S presents a reason for performing the action described in N.

- **Effect:** R's desire to perform action in N is increased.

- **Typical connectives:** Causal connectives.

- **Example:** [Running the Olympic Games are a win-win situation for any city.]$_S$ [Berlin must now apply for the 2026 games!]$_N$

- **Remark:** Sometimes W metaphorically encourages all Rs to do a certain action, as in the example given here.

Enablement

- **N:** An action to be performed by R.

- **N/S:** Comprehending S makes it easier for R to perform the action described in N.

- **Effect:** R's potential ability to perform the action in N increases.

- **Typical connectives:**

- **Example:** [Replace the spark plugs.]$_N$ [A square key wrench can be found right under the cap.]$_S$

- **Remark:** In the genre of editorial, this relation is rare.

4.3 Primarily semantic relations

Circumstance

- **S:** An event or state that actually occurs or has occurred (i.e., not a hypothetical one).

- **N/S:** S characterizes a framework in which N is to be interpreted, such as its temporal or locative position.

- **Effect:** R recognizes that S provides the framework for interpreting N.

- **Typical connectives:** *as; when; while; …* for a temporal frame.

- **Example:** [When Veag came under pressure because of the deregulation of the electricity market,]$_S$ [they compensated for this by squeezing their suppliers.]$_N$ (maz-5297, PCC)

CONDITION

- **S:** A hypothetical, future, or in other ways unreal situation.

- **N/S:** The realization of N depends on the realization of S.

- **Effect:** R recognizes how the realization of N depends on the realization of S.

- **Typical connectives:** *if .. then; in case; ...*

- **Example:** [When more and more communities are in financial trouble,]$_S$ [regularly replacing the school textbooks becomes just impossible.]$_N$ (maz-00002, PCC)
  [If the sanitary facilities are still not available in the coming season,]$_S$ [Radewege is in danger of losing the competition for attracting boats to the campground.]$_N$ (maz-6488, PCC)

OTHERWISE

- **N:** A hypothetical, future, or in other ways unreal situation.

- **S:** A hypothetical, future, or in other ways unreal situation.

- **N/S:** The realization of N impedes the realization of S.

- **Effect:** R recognizes the dependency relation of prevention between the realization of N and the realization of S.

- **Typical connectives:** *otherwise; ...*

- **Example:** [The city would be smart not to respond to this moral imperative.]$_N$ [Otherwise it would have to install speed bumps in front of every single school.]$_S$ (maz-5709, PCC)
  [The community council should revoke its decision,]$_N$ [before citizens start buying torches.]$_S$ (maz-15609, PCC)

UNLESS

- **N:** A hypothetical, future, or in other ways unrealized situation.

- **S:** A hypothetical, future, or in other ways unrealized situation.

- **N/S:** S determines the realization of N: N is only being realized if S is not being realized.

- **Effect:** R recognizes that N is realized provided that S is not realized.

- **Typical connectives:** *unless; ...*

- **Example:** [Tomorrow the satellite will fall into the Pacific Ocean.]$_N$ [Unless it burns up upon entering the earth atmosphere, after all.]$_S$

ELABORATION

- **N/S:** S provides details or more information on the state of affairs described in N (but not on a single entity mentioned in N; see E-ELABORATION below). N precedes S in the text. Typical relations between N and S are set::element, whole::part, abstraction::instance, procedure::step.

- **Effect:** R recognizes S as providing additional detail for the state of affairs in N. R identifies the element of subject matter for which detail is provided.

- **Typical connectives:** *in particular; for example; …*

- **Example:** [Diepensee will relocate.]$_N$ [No question about that.]$_S$ (maz-6993, PCC)

E-ELABORATION

- **N/S:** S provides details or more information on a single entity mentioned in N. N precedes S in the text.

- **Effect:** R recognizes S as providing additional detail for an entitiy in N. R identifies the element of subject matter for which detail is provided.

- **Typical connectives:** Rarely signalled by connectives.

- **Example:** [Today thousands of visitors want to sense the atmosphere of a historical classroom,]$_N$ [as it exists only in very few places in Germany.]$_S$ (maz-6728, PCC)

- **Remark:** In the example, the *as*-clause (S) does not provide information on the 'sense' activity, but on the 'classroom', i.e., an entity mentioned in N.

INTERPRETATION

- **N/S:** S shifts the content of N to a different conceptual frame. This does not imply an evaluation of N (or the evaluation is of only secondary imprtance). N precedes S in the text.

- **Effect:** R recognizes that S relates N to a framework of ideas not involved in the knowledge presented in N itself.

- **Typical connectives:** *thus; ...*

- **Example:** [Now for the first time we bought a Cadillac.]$_N$ [That probably makes us part of the bourgeoisie.]$_S$

- **Remark:** Whenever S primarily provides an assessment (on the positive/negative scale) of N, Evaluation is to be used, rather than Interpretation.

MEANS

- **N:** An activity or action.

- **N/S:** S provides information that makes the realization/execution of N more probable or simple (e.g., an instrument).

- **Effect:** R recognizes that the method or instrument in S tends to make realization of N more likely.

- **Typical connectives:** *thus; ...*

- **Example:** [In August, Berliners always enjoy travelling to Lichtenrade.]$_N$ [To that end, they usually take the S25 train.]$_S$
[The chairman of the party made sure to be on bad terms with two mayors]$_S$ [and thus will lose credibility and standing.]$_N$ (maz-18160, PCC)

CAUSE

- **N:** A state or event in the world.

- **S:** A state or event in the world.

- **N/S:** The state/event in N is being caused by the state/event in S.

- **Effect:** R recognizes S as a cause of N.

- **Typical connectives:** *because; since; therefore; ...*

- **Example:** *Unexpected press report on relocating the village of Diepensee.* [Mayer Jochen Wagner reacted with surprise, too.]$_N$ [After all, just on Monday the community council had agreed to expand the village of Diepensee.]$_S$ (maz-6993, PCC)

Result

- **N:** A state or event in the world.

- **S:** A state or event in the world.

- **N/S:** The state/event in S is being caused by the state/event in N.

- **Effect:** R recognizes N as a cause of S.

- **Typical connectives:** *because; since; therefore; ...*

- **Remark:** This relation is parallel to Cause. Deciding between the two depends solely on judging the relative importance of the segments for the text.

Purpose

- **N:** An activity or action.

- **S:** A hypothetical, future, or in other ways unrealized situation.

- **N/S:** S is being realized through the realization/execution of N.

- **Effect:** R recognizes that the activity in N is initiated in order to realize S.

- **Typical connectives:** *in order to; to; ...*

- **Example:** [In order to protect their troops,]$_S$ [the U.S. do not comment on news reports on the start of a ground operation in Afghanistan.]$_N$ (maz-5701, PCC)

- **Remark:** There is a causal relationship in a wide sense. The difference to the relations Cause/Result is that with Purpose, S is clearly marked as hypothetical/unrealized, and represents the intention or goal of the acting person.

Solutionhood

- **S:** The content of S can be regarded as a problem.

- **N/S:** The content of N can be regarded as a solution to the problem in N. N usually precedes S in the text.

- **Effect:** R recognizes N as a solution to the problem presented in S.

- **Typical connectives:** Rarely signalled by connectives.

- **Example:** [With the anti-smoker regulations being passed, many pubs will be caught in a trap.]*S* [They should start looking into possibilities for having separate rooms.]*N*

4.4   Textual relations

The following three relations primarily serve to organize the text and its understanding, so they belong neither to the semantic (describing the world) nor to the pragmatic (change the attitude of readers) group.

PREPARATION

- **N/S:** S precedes N in the text. S orients R toward the topic of N.

- **Effect:** R is more ready, interested or oriented for reading N.

- **Typical connectives:** Rarely signalled by connectives; sometimes by a colon.

- **Example:** [One thing is evident:]*S* [With each day of air raids, the military operations of the U.S. lose more and more credibility.]*N* (maz-5701, PCC)

- **Remark:** This relation is to be used when S does not serve any stronger purpose than setting the topic for N, or when it consists of an introductory formula. S should contain only minimal information on its own.

RESTATEMENT

- **N/S:** N precedes S in the text. S repeats the information given in N, using a different wording. N and S are of roughly equal size.

- **Effect:** R recognizes S as a restatement of N.

- **Typical connectives:** *in other words; ...*

- **Example:** [The mayor gave all the information to the councillors,]*N* [kind of filling them in completely.]*S*

SUMMARY

- **N:** N consists of more than one EDU.

- **N/S:** S succeeds N in the text and repeats the information given in N, but in a shorter form.

- **Effect:** R recognizes S as a shorter restatement of N.

- **Typical connectives:** *in short; ...*

4.5   Multinuclear relations

Among the multinuclear relations, Sequence belongs to the semantic type, while the others can work as semantic, pragmatic or textual, depending on context.

CONTRAST

- **N:** Exactly two nuclei. Both are of equal importance for W's purposes. The contents are comparable yet not identical. They differ in aspects that are important to W.

- **Effect:** R recognizes the comparability and the difference(s) yielded by the comparison being made.

- **Typical connectives:** *on the other hand; yet; but; ...*

- **Example:** [My first car was small.]$_N$ [The second was already a sizable limousine.]$_N$

SEQUENCE

- **N:** The nuclei describe states of affairs that occur in a particular temporal order.

- **Effect:** R recognizes the succession relationships among the nuclei.

- **Typical connectives:** *then; before; afterwards; ...*

- **Example:** [At nine o'clock the teacher entered the classroom.]$_N$ [Five minutes later she announced that a test will be written.]$_N$

- **Remark:** The states of affairs can be presented in their actual temporal order ("afterwards") or in the opposite one ("before that").

LIST

- **N:** The nuclei provide information that can be recognized as related, enumerating. They all contribute to the text function in the same way.

- **Effect:** R recognizes the comparability of linked items.

- **Typical connectives:** Comma, enumeration, ...

- **Example:** What I did yesterday: [Cook dinner,]$_N$ [look after the kids,]$_N$ [clean the bathromm.]$_N$

- **Remark:** See Section 4.6 for hints on handling elliptical contexts (as in the example).

CONJUNCTION

- **N:** The nuclei provide information that can be recognized as related, enumerating. They all contribute to the text function in the same way, and they are linked by coordinating conjunctions.

- **Effect:** R recognizes that the linked items are conjoined.

- **Typical connectives:** *and; or; ...*

- **Example:** [Should the aerial warfare have been meant merely as a menace,]$_S$ [it would lose its legitimacy at the latest when winter kicks in and triggers the inevitable human disaster.]$_N$ [And the U.S. would lose their backing in Western Europe.]$_N$]$_N$ (maz-2318, PCC)

- **Remark:** The functions of List and Conjunction are identical. When the surface condition for Conjunction holds, this relation is to be used.

JOINT

- **N:** The nuclei provide different kinds of information, which are not of the same type; yet they are not in a clearly identifiable semantic or pragmatic relation, nor do they form an enumeration. Still, there is a coherent link, as they contribute in similar ways to the overall text function.

- **Effect:** R recognizes that each nucleus contributes its own message, which however serve the same overall text purpose.

- **Typical connectives:** Additive connectives such as *in addition; also.*

- **Example:** See Section 6.

- **Remark:** Joint is to be used when a multinuclear relation is needed (from the text-global perspective) but none of the specific relations are applicable.

4.6 Remarks on relation assignment

In many cases, deciding on a relation can be difficult because the context produces ambiguity: multiple relations seem to be equally applicable. Also, decisions on recursive embedding (i.e., building hierarchical structure) are not always simple. In the following, we briefly discuss these problems.

4.6.1 Hierarchical structure

(1) The handling of **elliptical constructions** is sometimes tricky.

(22)   Nobody knows why she does not come, who takes her place, when or whether she will come back. (maz-5876, PCC)

The modal embedding *nobody knows* has scope over the entire sentence, which is to be divided into different segments in case there are clear hints for a coherence relation to be present. In this sentence, the enumeration of propositions serves as such a hint. The criterion for segmentation is that a proposition is by and large complete. That holds for *why she does not come* and for *who takes her place*, but not for the highly ellipctical *when* – the individual pronoun is not a complete segment. Therefore, the analysis results in a LIST relation with three segments: (Nobody knows why she does not come), (who takes her place), (when or whether she will come back).

(2) Sentence-final **punctuation** in most cases takes priority for building hierarchical structure—but not always. In the following example, the nucleus of the CONDITION starts with *it would* and ends with *Western Europe*; thus the two CONJUNCTION nuclei are embedded in the CONDITION nucleus:

(23)   [Should the aerial warfare have been meant merely as a menace,]$_S$ [[it would lose its legitimacy at the latest when winter kicks in and triggers the inevitable human disaster.]$_N$ [And the U.S. would lose their backing in Western Europe.]$_N$]$_N$

(3) A context is **ambiguous** for hierarchical structure when, for example, a claim is being supported by more than one subsequent argument. These can either be linked individually to the claim, or they can be first conjoined by a LIST or CONJUNCTION relation, so that a single supporting segment results. In such cases, we use the latter option: Segments playing the same role in context should form one complex segment in a LIST or CONJUNCTION.

4.6.2 Editorial genre: Make nuclearity decisions carefully

These guidelines are to some extent tailored to the annotation of editorials. These are opinionated texts with the goal of suggesting a particular opinion to the reader. This makes the role of an RST tree—to represent the communicative purposes of the writer—particularly important. In order to achieve a good reconstruction of the writer's plan, we have to be careful when deciding on

the nuclearity status of segments on any level of the tree. For these decisions, one should pay attention to the *strong nuclearity principle* proposed by Marcu (2000). It states that the nucleus status of a segment is propagated throughout the tree, which is relevant both for the *bottom up* and the *top down* reading: A segment that is considered as central for the text should be nuclear not only when linking it to its immediate neighbours, but the resulting larger segment should likewise be nuclear in its own context, and so forth. Seen the other way round, this means that a relation that holds between two large segments should in particular apply to the nuclei of these segments.

Occasionally, however, authors of editorials produce a less connected sequence of points that seem equally important, so that any differentiation in terms of nuclearity would be arbitrary. In these cases, a multinuclear relation is to be preferred.

4.6.3 Editorial genre: Prefer pragmatic relations

Along the lines of the goal of reconstructing the writer's communicative plan, in our genre it makes sense to give a general preference to pragmatic relations, as these serve specifically to represent configurations of the writer's intentions. Thus, when at some point during an analysis there is an ambiguity between a pragmatic and a semantic relation, the former is to be preferred, as it yields a more important contribution to the reconstructed plan.

4.6.4 Prefer more informative over less informative relations

Our final principle can help in particular with making a decision among the semantic relations. Their definitions include different degrees of "informativeness" in terms of the implicit meaning that is assigned to them. For example, ELABORATION merely states that the topic of the discourse is being continued in a more specific way. In a context where that is the case, but in addition there is, e.g., a causal connection between the segments, then a CAUSE/RESULT relation is to be preferred. In a similar way we can characterize the differences between some similar semantic or pragmatic relations. Consider CIRCUMSTANCE, which is not very informative: It only states that that one segment is a framework in which the other is to be interpreted. The applicability of this relation can easily overlap with that of BACKGROUND; the latter, however, focuses on the effect of easier understanding of the nucleus. This is a stronger, more informative criterion, and therefore BACKGROUND is to be preferred over CIRCUMSTANCE in such situations.

4.6.5 Differences between similar relations

Adversativity: The first thing to consider when deciding among the similar relations CONTRAST, ANTITHESIS and CONCESSION is multinuclearity. If neither segment is deemed more important than the other, then CONTRAST is to be chosen.

(24) [My sister loves cats.]$_N$ [My brother, on the other hand, is keen on dogs.]$_N$

Between the two mononuclear relations, Concession[4] is the more specific one; it generally involves a violated or failed expectation. Concession is often used to convince the reader of a particular point by increasing the reader's regard towards N:

(25) [Although you are correct that the app is cheap]$_S$ [I nevertheless wouldn't buy it, because it has many bugs.]$_N$

Sometimes, Concession is used in order to prevent false implicature (i.e., an inference that logically emerges, but is not implied) presented in the S-N order. Here, the writer knows that the reader might draw an inference from a fact stated in S, and the inference is based on general world knowledge; but the writer wants to prevent such inference. For example:

(26) [The classrooms are small,]$_S$ [though they are not unsuitable.]$_N$

Concession may also represent an unexpected, surprising move towards what is being said from what was said before. In this sense, Concession can also be thought in terms of a negative-causal relation in which an expected causal relationship does not hold—a cause in S does not have the consequence one would anticipate from a law of cause and effect. For example:

(27) [Although it was December,]$_S$ [no snow fell and the temperature rose to 20 degrees.]$_N$

On the other hand, in an Antithesis relation, the writer assigns different or unequal weights to the propositions which are usually presented in S-N order. The writer often dismisses S in order to establish or reinforce N. For example:

(28) [Annuities are rarely a good idea at age 35 because of the withdrawal restrictions.]$_S$ [But at age 55, "they may be a great deal," says Mr. Wilson, the Columbia, S.C., planner.]$_N$ (wsj_0689, RST-DT)

One important difference between Antithesis and Concession is that the claim which is represented by S is dismissed in Antithesis, but is acknowledged in Concession.

Sometimes, Antithesis is equated with correction because in both cases the second segment serves as a complete substitute for the initial segment in meaning or implication. However, unlike in correction, where the first proposition is fully valid until the correction marker is produced, in Antithesis the substitution is foreshadowed, because the first proposition is marked as invalid from the very beginning (indicated by syntactic negation, with *not*, *n't*, or lexically). Antithesis of this kind includes examples such as the following:

(29) [This is not coke,]$_S$ [this is red wine.]$_N$

---

[4]For a more extensive discussion of this relation, see Grote et al. (1997). Some of our discussion here is based on parts of that paper.

For the purpose of distinguishing Concession from Antithesis, paraphrasing the segment pair with an *although* construction (if the conjunction is not overtly present already) is usually a good indicator: If it yields essentially the same meaning, then Concession is the right relation; otherwise, the relation Antithesis is more appropriate, which operates more on the level of evaluation of the segments' content.

Finally, here are some more corpus examples to help differentiate between Antithesis and Concession:

(30) Concession: [Although all the price data were adjusted for normal seasonal fluctuations,]$_S$ [car prices rose beyond the customary autumn increase.]$_N$ (wsj_2358, RST-DT)

(31) Concession: [Monsanto Co., too, is expected to continue reporting higher profit,]$_N$ [even though its sales of crop chemicals were hurt in the latest quarter by drought in northern Europe and the western U.S.]$_S$ (wsj_2398, RST-DT)

(32) Concession: [A severe recession could, of course, raise delinquency rates,]$_S$ [but so far the current levels of consumer debt don't seem to loom as a major threat.]$_N$ (wsj_1389, RST-DT)

(33) Antithesis: [Annuities are rarely a good idea at age 35 because of the withdrawal restrictions.]$_S$ [But at age 55, "they may be a great deal," says Mr. Wilson, the Columbia, S.C., planner.]$_N$ (wsj_0689, RST-DT)

(34) Antithesis: [The Northern California earthquake and Hurricane Hugo are likely to temporarily damp sales growth in the West and South Carolina.]$_S$ [But Mr. Lagnado predicted the regional trends would continue through Christmas.]$_N$ (wsj_0640, RST-DT)

(35) Antithesis: [The merger must be formally approved by the partners of both firms]$_S$ [but is expected to be completed by year end.]$_N$ (wsj_0617, RST-DT)

Causality: The field of causality or justification comprises quite a few relations. First, the basic decision between primarily semantic (report on event causality in the world) and primarily pragmatic (justification as part of an argument, aiming to change the reader's mind) is to be applied. Among the semantic relations, Condition, Otherwise, Unless and Purpose are restricted for potential, hypothetical, or future—hence *non-factive*—connections between cause and effect. ("If the sun shines tomorrow, we'll do a boat trip"). Here, Purpose has a specifically goal-oriented facet: The nucleus is an activity that is performed for achieving the state of affairs in the satellite. The differences between Condition, Otherwise and Unless result from the particular role of negation, as described in their definitions. But if cause and effect are being described as *factive*, then Cause or Result are to be used, which differ solely in the distribution of nucleus and satellite.

In the set of pragmatic relations, Motivation can only be used when the reader is encouraged to perform a certain activity (nucleus), on the grounds of

the satellite. EVIDENCE/REASON/JUSTIFY all want to change the attitude of the reader, and they differ in terms of the type of justification: With EVIDENCE, the claim is supported by information that is presented as 'objective', such as an observation from the world. ("party X has lost half of its membership → party X will lose the upcoming election"). With REASON, on the other hand, the support is also a subjective assessment or thesis given by the writer ("party X will not have a chance at the upcoming election → One should not vote for party X"). JUSTIFY is similar, but here the support is from stating a basic attitude of the writer, rather than a specific thesis. This attitude can be a political orientation, a moral value, etc.

Additive relations (LIST/CONJUNCTION/JOINT): These three relations have similar ranges of application. For LIST and CONJUNCTION, the parallelism of the functions of the segments is of central importance: They have to play the same role for the intention of the segment they are being attached to. A proto-typical case is an enumeration of statements that all serve to illustrate the same general idea. CONJUNCTION is to be used when a lexical signal (a connective such as *and*, *in addition*, etc.) signals the relation. This can be within a single sentence, or between separate ones. If there is no connective, LIST is to be used. For JOINT, the parallelism constraint does not hold. This relation applies whenever the author makes two or more points that are juxtaposed in a relatively loose way, and which do not have the same function for the higher-level segment. Furthermore, all points have to be equally important—otherwise, a suitable mononuclear relation is to be chosen. In practice, JOINT is not very frequent, and usually it applies only for connecting larger segments.

For further illustration, the sample text that we discuss at the end of the guidelines has instances of all three relations.

5 Annotation procedure

Start with the annotation only after you read the text completely and understood its line of argumentation. Then, build the analysis following the steps described below.

Note: The headline of the text constitutes an EDU but is not to be integrated into the tree; it just remains as an isolated, "artificial" segment.

1. Is the text composed of recognizable topical units? Mark the boundaries between units that deal with different aspects of the topic (if any). These boundaries will later delimit larger text segments in the RST tree.

2. Select the EDUs that play an important role for the text. At the end of the analysis, these should end up as "strongly nuclear". If one EDU can be singled out as representing the central claim of the text, mark it as such, and also the other important EDUs. You can check the result of this step with a "paraphrase test": Put all the marked EDUs in a sequence (deleting any dangling connectives, and replacing anaphors with full NPs)

and judge whether the resulting short text is an adequate summary of the original. If necessary, revise your choice of important nuclei.

3. Going left to right through the text, consider each EDU and its direct neighbours. Is there a clearly recognizable relation between such a pair? This will often be the case with syntactically dependent pairs, and sometimes when two independent EDUs are linked with a connective.
   For each pair of EDUs that are to be joined:

   - Decide whether one of the EDUs is more important than the other, or whether both are of equal weight.
   - Decide which coherence relation holds between the two EDUs. This is constrained by the nuclearity decision you already made.
   - In case you choose a multinuclear relation, more than two EDUs might belong together; check this.

4. When all pairs of neighbouring EDUs have been checked, continue by considering the larger units. A connective can join longer units than a single EDU, and of course, relations between EDUs and/or larger units can also be unsignalled. First, mark the relations that are easy to identify. If you established topic boundaries in Step 1, consider these as relations between suitable segments. When choosing nuclei, the role of the nucleus segment for the whole text can now pay a role: Your decisions on assigning relations between larger segments should make sure that in the end, the EDUs selected in Step 2 are strongly-nuclear. For deciding on a relation, it often helps to check whether a prototypical connective (see the definitions of the relations) can be inserted. In general, at this stage, be prepared to revise earlier decisions—an analysis takes time and often will involve weighing decisions in the light of others that are taken later on when the tree structure becomes clearer.

5. In marking the relations between larger segments, it is advisable to proceed in a *bottom-up* fashion: Conjoin EDUs and/or neighbouring larger segments, and successively construct the tree moving upward.

6   Analysis of a sample text

In order to illustrate the annotation process, we present a sample text (provided in Fig. 3) and explain the development of its RST structure following our guidelines. The numbers in the text represent EDUs.

Thematic boundaries: As Figure 4 shows, the text comprises two broad topic units, represented by the spans 1-9 and 10-14. Span 1-9 introduces the topic of DPC Acquisition Partners' petition filing against Dataproducts Corp. It also projects about the probable outcome of the petition, and reports on the responses made by the two business groups on the petition. On the other hand,

(1) DPC Acquisition Partners, a hostile suitor for Dataproducts Corp., filed a petition in federal district court in Los Angeles (2) seeking to have its standstill agreement with the computer printer maker declared void, (3) and it proceeded with a \$10-a-share tender offer for the company.

(4) The offer would give the transaction an indicated value of \$189 million, (5) based on the 18.9 million shares the group doesn't already own. (6) DPC holds about 7.8% of Dataproducts' shares.

(7) Lawyers representing DPC declined to elaborate, (8) saying they didn't have a final copy of the filing. (9) Jack Davis, Dataproducts' chairman, said he hadn't yet seen the filing and couldn't comment.

(10) DPC made a \$15-a-share bid for the company in May, (11) but Dataproducts management considered the \$283.7 million proposal unacceptable. (12) Dataproducts had sought a buyer for several months, (13) but it is now in the midst of a restructuring plan and (14) management says the company is no longer for sale.

Figure 3: Sample text wsj-0697 from the RST Discourse Treebank

span 10-14 reports on the present status on Dataproducts Corp., while making a reference to its past sale-out ventures.

Nuclearity at the text level: Within the text, span 1-9, span 1 and span 3 play an important role. They document two successive events: the petition filing by DPC Acquisition Partners, and their subsequent move with a tender offer for their rival company. These two spans establish one of the main themes of the text. Within span 10-14, on the other hand, span 13 and span 14 serve an important function. They present Dataproducts Corp.'s status regarding its ongoing internal restructuring and the management's recent take on its (anticipated) sale. These two spans constitute the other central theme of the text.

Local EDU links: At the most local level, we find the following spans are related: span 1 and span 2; span 4 and span 5; span 7 and span 8; and finally, span 10 and span 11. Span 2 (satellite) is a present participial clause, and it enters into a Purpose relation with span 1 (nucleus). The Purpose relation is signalled by the word *seeking* in span 2. Span 4 (nucleus) and span 5 (satellite) are connected by a Background relation. The relation is inferred from the use of the past participial clause (acting as span 5) and also by the word *based* in span 5. Span 8 (satellite) is linked with span 7 (nucleus) by a Reason relation which is based on a lexical chain of words/phrases across the spans (declined to elaborate — didn't have the final copy). Finally, a multinuclear Contrast relation exist between span 10 and span 11, and the relation is indicated by the connective *but*.

Higher units and hierarchization: Looking beyond the local level of the RST analysis, we build the hierarchical structure involving the larger units in the text. First, span 1-2 (nucleus) and span 3 (nucleus) present two successive events, and thus, are connected by a multinuclear SEQUENCE relation which is conveyed in a straightforward (albeit underspecified) way by the connective *and*. Then, span 4-5 (satellite) projects on the probable outcome of the moves made by DPC represented in span 1-3 (nucleus). Thus, these spans are related by an INTERPRETATION relation. Span 6 states about the percentage of Dataproduct's shares owned by DPC. This serves as a kind of background information, and helps the readers understand the events and their probable outcomes more clearly. Thus, span 6 (satellite) is linked with span 1-5 by means of a BACKGROUND relation. Span 7-8 and span 9 list the official responses from DPC and Dataproduct, respectively, on the events presented in span 1-3. Accordingly, they act as nuclei and enter into a LIST relation. Furthermore, since the span 7-9 (satellite) provides additional information about the events in span 1-6 (reporting on the customary comments typically made by the spokespersons of companies on issues which concern them), it relates to span 1-6 (nucleus) by an ELABORATION relation.

Looking forward through the analysis, span 10-11 reports on DPC's proposal for buying Dataproduct's shares in the past and also Dataproduct's negative stand in response to the offer. Span 12 mentions about Dataproducts' previous sale-out ventures for months. Span 10-11 and span 12 thus do not form an itemized list, yet they are somehow related by a semantic relation which is hard to specify. That is why these spans, functioning as nuclei, enter into a JOINT relation with each other. Next, span 10-12 and span 13 describe facts which are set against different times (signalled through the connective *but* and *now*). The interpretation of span 13 (Dataproduct's current status in terms of its ongoing internal restructuring) is facilitated by the background information about the events in span 10-12 (regarding the company's sale-out in the past). Accordingly, span 10-12 (satellite) relates to span 13 by a BACKGROUND relation. Then, span 14 reports on what Dataproduct's management says about its present plan for sale-out. Thus, span 10-13 and span 14 are related by the same topic, and they are linked by a multinuclear CONJUNCTION relation marked by the conjunction *and*.

Finally, the last decision concerns the connection between span 1-9 and span 10-14. Span 1-9 reports on DPC's recent buy-out moves towards Dataproducts, and span 10-14 reports on Dataproduct's past sale ventures and its present stand against the sale. These topics are related semantically, although they do not form an enumerated list nor are conjoined by typical conjunctions. That is why, they function as nuclei and connect to each other by a JOINT relation.

Figure 4 shows the RST tree resulting from these decisions.

Figure 4: RST analysis for sample text wsj-0697

References

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

Jyotibhusan Chaki. *Bangla Bhashar Byakaran*. Ananda Publishers, Kolkata, India, 1996.

Udaya Kumar Chakraborty. *Bangla Padaguccher Sangathan (Structure of Bengali Phrases)*. Dey's Publishing, Kolkata, India, 1992.

Suniti Kumar Chatterji. *Bhasha-Prakash Bangla Vyakaran*. Rupa and Company, New Delhi, India, 1988.

Debopam Das and Manfred Stede. Developing the Bangla RST Discourse Treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, to appear.

Debopam Das, Maite Taboada, and Manfred Stede. The good, the bad, and the disagreement: Complex ground truth in rhetorical structure analysis. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 11–19, 2017.

Brigitte Grote, Nils Lenke, and Manfred Stede. Ma(r)king concessions in English and German. *Discourse Processes*, 24(1):87–118, 1997.

William Mann and Sandra Thompson. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text*, 8:243–281, 1988.

Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA, 2000.

Pabitra Sarkar. *Bangla Byakaran Prasanga*. Dey's Publishing, Kolkata, India, 2006.

Manfred Stede. *Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0*. Universitätsverlag Potsdam, 2016.

Manfred Stede and Arne Neumann. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 925–929, Reikjavik, 2014.

Maite Taboada and William Mann. Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, 8(4):423–459, 2006.

Milan Tofiloski, Julian Brooke, and Maite Taboada. A syntactic and lexical-based discourse segmenter. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 77–80, Singapore, 2009.