Discourse Segmentation in Bangla

Debopam Das
University of Potsdam
Karl-Liebknecht Strasse 24-25,
14476 Potsdam, Germany
debdas@uni-potsdam.de

Abstract

An important kind of discourse annotation is relational annotation in which texts are analyzed with respect to coherence relations (relations between text components, such as *Cause* or *Evidence*) present in the texts. Relational annotation according to Rhetorical Structure Theory (Mann and Thompson, 1988) typically begins with segmenting a text into minimal discourse units, which are then linked with each other (and later recursively with larger units) by certain coherence relations. As part of an ongoing corpus development project called the Bangla RST Discourse Treebank (Das and Stede, to appear), we have considered, examined and implemented a number of segmentation principles and strategies for dividing Bangla texts into minimal discourse units for the purpose of relational annotation. In this paper, we provide an overview of our annotation tasks, and describe our segmentation guidelines. We also present a few problems we encountered in segmenting Bangla texts, and discuss how we have addressed those issues.

Keywords: Bangla RST Discourse Treebank, discourse segmentation, Rhetorical Structure Theory, Bangla

1. Introduction

Relational annotation is a kind of discourse annotation that provides analysis of a text with respect to coherence relations (*Cause*, *Elaboration* or *Evidence*) that hold between the text components. Relational annotation tasks, according to Rhetorical Structure Theory or RST (Mann and Thompson, 1988), as followed in a number of RST-based discourse corpora, usually involves a number of sequential steps, typically beginning with the segmentation of texts into minimal discourse units. In RST, clauses are generally considered to be the basic units of discourse (Tofiloski et al., 2009). Nevertheless, RST segmentation policies differ from studies to studies, primarily because clauses are treated in different ways as information-bearing units, and partly because exceptions in the text data are handled in various manners.

We deal with segmentation of texts as part of an ongoing corpus development project called the Bangla RST Discourse Treebank or Bangla RST-DT (Das and Stede, to appear). This project builds a discourse corpus in Bangla which is annotated for coherence relations. RST-based corpora have been created for English (Carlson et al., 2002) and many other European languages, such as German (Stede, 2016), Dutch (van der Vliet et al., 2011), Brazilian Portuguese (Cardoso et al., 2011), Spanish (da Cunha et al., 2011) and Basque (Iruskieta et al., 2013). The practice has also been expanded to corpora in Asian languages such as Chinese (Cao et al., 2017) and Russian (Toldova et al., 2017), which are currently under production. We decide to contribute to this tradition by developing an RST corpus in Bangla, which, to our knowledge, is going to be the first dataset of its kind. As part of the relational annotation tasks, we have considered, examined and implemented a number of segmentation principles and strategies for dividing Bangla texts into minimal discourse units. In this paper, we present our segmentation guidelines, and discuss a few challenges associated with segmenting Bangla texts.

This paper is organized as follows: In Section 2., we provide a brief introduction of coherence relations and RST. Section 3. presents an overview of the Bangla RST-DT. In Section 4., we state the theoretical underpinnings of our segmentation guidelines, and describe different segmentation principles followed in the annotation. Section 5. presents a few issues in segmenting Bangla texts, and discusses how we have addressed them. Finally, Section 6. summarizes the paper, and provides the conclusion.

2. Coherence Relations and RST

The concept of coherence relations has been extensively studied in different discourse theories (see Das and Stede (to appear) for a list of theories and references), among which we chose to use Rhetorical Structure Theory or RST (Mann and Thompson, 1988) for our relational annotation purpose. This is because we believe that certain aspects of text organization are best captured by RST. We also chose RST because it is essentially a language neutral theory and it has been successfully used in many computational applications, such as text generation, discourse parsing, and text summarization (see Taboada and Mann (2006) for an overview).

Text organization in RST is described in terms of relations that hold between two or more non-overlapping text spans (discourse components). Relations can be multinuclear, reflecting a paratactic relationship, or nucleus-satellite, a hypotactic type of relation. The names nucleus and satellite refer to the relative importance of each of the relation components. Relation inventories are open, but the most common ones include names such as *Cause, Concession, Condition, Elaboration, Result* or *Summary*.

Texts, according to RST, consist of basic discourse units (also called elementary units or EDUs) that are connected to each other (or to larger units comprising two or more

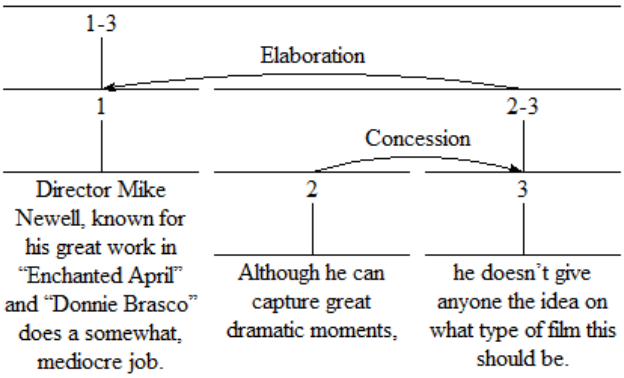

Figure 1: Graphical representation of an RST analysis

EDUs) by rhetorical (or coherence) relations in a recursive manner. According to Mann and Thompson (1988), the recursive application of different types of relations can be used to capture the entire structure of most texts. This, in practice, means that the RST analysis can be developed and represented as a tree structure in which the clausal units stand for the branches and the relations stand for the nodes.

For the purpose of illustration, we provide the annotation of a short text[1], represented by the tree diagram[2] in Figure 1. The text is segmented for three EDUs (minimal spans), which are marked by the cardinal numbers 1, 2 and 3, respectively. In the diagram, the arrow points to a span called the nucleus, and away from another span called the satellite. Span 2 (satellite) is connected to Span 3 (nucleus) by a *Concession* relation, and together they make the combined Span 2-3, which is further linked as a satellite to Span 1 (nucleus) by an *Elaboration* relation.

3. Bangla RST Discourse Treebank

Bangla RST-DT (Das and Stede, to appear) is a corpus of Bangla (currently under production) which is annotated for coherence relations following RST. The corpus contains 266 texts, comprising 71,009 words, with an average of 267 words per text. The corpus represents the newspaper genre. The texts have been collected from a popular Bangla daily called *Anandabazar Patrika* published in India. The texts in the corpus come from eight different sub-genres: (1) business-related news, (2) editorial columns, (3) international affairs, (4) cityscape (stories on Kolkata, the home city of the newspaper), (5) letters to the editor, (6) articles on nature, (7) features on science, and (8) reports on sports.

The annotation guidelines followed in the corpus[3] are based on the guidelines previously used in the Potsdam Commentary Corpus or PCC (Stede, 2016)[4], and are more closely related to an updated version of the PCC guidelines used in (Das et al., 2017)[5]. The corpus employs a set of 31 RST relations (26 mononuclear and 5 multinuclear relations), which are further divided in three groups: semantic, pragmatic and textual relations.

The Bangla RST-DT started with the annotation of 16 texts, taking two texts from each of the eight sub-genres mentioned above. The texts were pre-segmented by an expert annotator (the author of the present paper), and then they were separately annotated by three (one expert and two trained) annotators who are all native speakers of Bangla. The annotations were evaluated for inter-annotator agreement, with respect to span determination, nuclearity status assignation and relation labeling. The scores showed fairly high level of agreement between annotators, which indicates that our annotations are reliable. The currently-ongoing work includes the annotation of the remaining 250 texts, and we expect to complete the production of the corpus within the next few years. For more information about the corpus, see Das and Stede (to appear).

4. Segmentation in Bangla RST-DT

RST-based discourse segmentation strategies have been implemented (although with a moderate range of variation) by many previous studies for different languages, such as English (Tofiloski et al., 2009; Carlson and Marcu, 2001), German (Lüngen et al., 2006; Sidarenka et al., 2015), Brazilian Portuguese (Pardo and Nunes, 2008), Dutch (Abelen et al., 1993; den Ouden et al., 1998; van der Vliet et al., 2011) and Basque (Iruskieta et al., 2013).

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). Both PCC and SLSeg guidelines closely adhere to the original definition of spans in RST, according to which clauses constitute EDUs containing a verb, either finite or non-finite. More particularly, only adjunct, and not complement clauses, form legitimate EDUs. Broadly, coordinated clauses (but not coordinated verb phrases), adjunct clauses and non-restrictive relative clauses are considered as EDUs.

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.

[1] Text source: SFU Review Corpus (Taboada, 2008)

[2] The RST diagram is created by RSTTool (O'Donnell, 2000) which provides a graphical representation of the RST analysis of a text in the form of a tree diagram. The tool is also used for doing the annotations in the Bangla RST-DT.

[3] http://angcl.ling.uni-potsdam.de/pdfs/Bangla-RST-DT-Annotation-Guidelines.pdf

[4] http://angcl.ling.uni-potsdam.de/resources/pcc.html

[5] http://www.sfu.ca/~mtaboada/docs/research/RST_Annotation_Guidelines.pdf

Although our segmentation guidelines are primarily meant to facilitate the annotation process in the Bangla RST-DT, the broader goal is to provide a set of RST-based discourse segmentation principles for Bangla, which can also be used for other Indo-Aryan languages, such as Assamese, Oriya or Punjabi. We believe that these guidelines can be adopted, modified and implemented according to specific annotation goals, and also that anyone having the basic knowledge of Bangla syntactic structures will be able to adequately follow them. Furthermore, since our segmentation principles mainly rely on formal criteria, they can also be used for the purpose of (semi-)automatic text segmentation, using the taggers and parsers available for Bangla (Hoque and Seddiqui, 2015; Ekbal and Bandyopadhyay, 2008; Hasan et al., 2010; Ghosh et al., 2009).

In the following subsection, we enumerate specific guidelines used for segmenting texts in the Bangla RST-DT. Most of the examples (accompanying specific guidelines) are taken from the corpus. 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).

4.1. Segmentation guidelines for Bangla

4.1.1. 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 absent on the surface, 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]

4.1.2. Pro-drop Constructions

Bangla is a pro-drop language, in which subject pronouns are omitted from clauses on many occasions. In our annotation, we consider such (adjunct) clauses (clauses only with verbal predicates, and not the overt subjects) as EDUs.

(2) [er par Ar pratiyogitAmulak
this.Gen after anymore competitive
Asare nAmben nA.]
tournament will participate not
(He) will not participate in competitive tournaments anymore after this. [sports-03]

4.1.3. 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
proper governmental provision be
jaruri.]
necessary
Having 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
such provision at all.Emph not that
sambhabata balA yAbe nA.]
probably say can not
That there is no such provision at all cannot be said. [editorial-column-08]

4.1.4. 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
many people doctors' to chamber go.Gen
cheYe jyotiShIr chembAre jAoYA beshi
than astrologers' to chamber go much
paChanda karen.]
prefer do
Many people prefer to go to astrologers' chambers than doctors' chambers. [letters-to-the-editor-06]

(6) [jiesTi kiChuTA hAsi phoTAte chaleChe
GST a little smile to bring go.Prog
bAik bhaktoder mukheo.]
motorcycle fans' face.Emph
GST is also going to bring a little smile on the faces motorcycle fans. [business-06]

4.1.5. Attribution Clauses

Attribution clauses are a kind of complement clauses, which 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 in a matrix clause, they are not assigned the status of EDUs.

(7) [praphesar AYAn hoYAT boleChen, "ekhonai
professor Ian Howat said now.Emph
Ata.mkita haYe parar konao kAron nei."]
panicked be get.Gen any.Emph reason not
Professor Ian Howat said, "There is no reason to get panicked by now." [science-04]

(8) [goYendApradhAn Aro jAnAn,
the chief of detectives more informed
dhritader jiGYAsAbAd karA hochChe.]
arrested ones' interrogation 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 t.NAr bA.Dii Chilo
of the attack primary target his house was
bale sandeha karChen tadantakArIrA.]
that suspicion do.Prog investigators
The investigators are suspecting that the primary target of the attack was his house. [international-01]

4.1.6. 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.

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

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

4.1.7. 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 *ye hetu / se hetu, yeman (yeman) / teman (teman), yadi / tabe,* etc. Clauses with such connectives are considered to be EDUs in our annotation.

(12) [... hAmlA ye hetu tIrthayAtrIder upar,]
... the attack since the pilgrims.Gen on
[se hetu ei hAmlAr ek anYatara
that is why this of the attack a different
tAtparya kh.NojAr chestA hachChe.]
significance find.Gen attempt being
Since the attack was on the pilgrims, that's why there is being an attempt to find a different significance of the attack. [editorial-column-07]

4.1.8. Nominal Modifiers

Nominal modifiers represented by verbal nouns are not considered as EDUs. In Example 13, the noun *'bAsTike'* ('the bus') is modified by the verbal noun *'ulTo dik theke AsA'* ('coming from the opposite side') and hence, it is not segmented as an EDU.

(13) [ulTo dik theke AsA bAsTike dhAkkA mAre
opposite side from come the bus hit
oi gA.DiTi.]
that car
The car hit the bus coming from the opposite side. [international-01]

4.1.9. Participial Clauses

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

(14) [dvitIYa TesTe phire ese] [sirij 1-1
second in the test coming back series 1-1
karlen phAph duplesi.]
did Faf du Plessis
Coming back in the second test, Faf du Plessis made the series 1-1. [sport-08]

4.1.10. Verbal Nouns with a Postposition

Verbal nouns, as already shown in Example 3 and 5, are not considered to be EDUs. However, when verbal nouns are used with a postposition, they are treated as EDUs. In Example 15, the verbal noun *'eman sambhAbanAder chine neoAr'* ('recognizing such potentials') with the postposition *'janya'* ('for') forms an EDU.

(15) [eman sambhAbanAder chine neoAr
such potentials recognize.Gen
janya] [upayukta sarkAri bandobasta thAkA
for proper governmental provision be
jaruri]
necessary
Having the proper governmental provision is necessary for recognizing such potentials. [editorial-column-08]

4.1.11. Infinitival Clauses

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

(16) [nyAnoke bhabiShyate rAstAy chAlAte]
Nano in the future on road run
[dubaCharer madhyei chAi natun lagni.]
of two years within.Emph want new investments
New investments are required within the next two years in order to run Nano on road in the future. [business-05]

4.1.12. Conditional Clauses

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

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

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

4.1.13. Coordinated Constructions

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

(18) [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.).

5. Segmentation Issues and Resolutions

For us, the biggest challenge was to perform the RST segmentation for a non-European language, for which no previous documented effort on discourse segmentation was available. In particular, we have encountered a few issues in our segmentation task, which are described below:

1. Bangla employs the use of phrasal verbs, which (unlike in English) comprise a pre-verbal element and the main verb (which is marked for tense and person). In certain instances, we have noticed that the phrasal verb constructions and adjunct clause pairs have similar forms, and it is often difficult to distinguish them. For instance, Example 19 and 20 are very similar in form. However, in Example 19 the form *khete* is a pre-verbal element of the phrasal verb *khete baseChen*, while in Example 20 *khete* acts as an infinitival adjunct clause (with the implication "in order to eat") (cf. (Chakraborty, 1992), p. 137-138).

   (19) tini khete baseChen.
   he/she to eat sat down
   He/she sat down to eat.

   (20) tini khete geChen.
   he/she to eat went
   He/she went to eat.

   For this problem, we use a paraphrase test: We checked whether it is possible to replace the questionable item *khete* ('to eat') with *khAbAr janya* ('for eating' or 'in order to eat'), and if the modified construction still yields a grammatical output, then we consider it to be an adjunct clause (and hence an EDU). We used this test and other similar tests for resolving such ambiguities.

2. Some texts in our corpus contain long speeches (whether direct quotes or indirect reported speeches). According to our guidelines for attribution clauses, we 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 21).

   (21) "[bhAloi haYeChe daurTA.] [Ami saThik pathei yAchCHi.] [tabe ekhanao anek kAj bAki.", baleChen bolT.]
   good.Emph has been the (sprint) race I right in-the-direction.Emph moving However still many things remaining said Bolt
   "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. Bangla makes use of correlatives (a pair of two particles) where one part presupposes the presence of the other. In the standard Bangla grammar (Chakraborty, 1992; Sarkar, 2006), correlatives provide a cover term for elements such as *yini / tini, yata / tata, ye hetu / se hetu, yeman / teman*, or *yadi / tabe* (see Section 4.1.6. and 4.1.7.). However, we have observed that these correlative elements have two distinct functions from a discourse point of view: Some correlatives (*yini / tini, yata / tata*, etc.) are used to establish coreferential relation between objects or entities, while others (*ye hetu / se hetu, yeman / teman, yadi / tabe*, etc.) are used for relating clauses or text spans. For this reason, we distinguish these two types in our annotation, and classify the former type as *correlative pronouns* (used in relative clauses) and the latter as *correlative discourse connectives* (used for linking clauses or text spans).

6. Conclusion

In this paper, we have presented the segmentation guidelines for annotating texts in the Bangla RST Discourse Treebank. We have discussed different segmentation principles and strategies, and motivated our reasons for choosing or developing those guidelines. Performing the segmentation for Bangla has also posed a few challenges for us, which we have successfully dealt with in our annotation task. We believe (as we have experienced) that in order to develop a set of RST segmentation guidelines in a new language one could adopt the basic segmentation principles from the available and recognized guidelines (such as the one for PCC or SLSeg), which could later be complemented by the language-specific guidelines or a modification of previous guidelines.

7. Acknowledgements

I would like to thank Dr. Manfred Stede for his active and sincere commitment to the Bangla Discourse Treebank project. Special thanks go to Dr. Pabitra Sarkar and Dr. Samir Sarkar for their invaluable suggestions on the topic of this paper.

8. Bibliographical References

Abelen, E., Redeker, G., and Thompson, S. (1993). The Rhetorical Structure of US-American and Dutch Fund-Raising Letters. *Text - Interdisciplinary Journal for the Study of Discourse*, 13(3):323–350.

Carlson, L. and Marcu, D. (2001). Discourse Tagging Manual. ISI Technical Report ISI-TR-545, University of Southern California.

Chaki, J. (1996). *Bangla Bhashar Byakaran*. Ananda Publishers, Kolkata, India.

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

Chatterji, S. K. (1988). *Bhasha-Prakash Bangla Vyakaran*. Rupa and Company, New Delhi, India.

Das, D., Taboada, M., and Stede, M. (2017). 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.

den Ouden, H. J. N., van Wijk, C. H., Terken, J. M., and Noordman, L. G. (1998). Reliability of discourse structure annotation. IPO Annual Progress Report.

Ekbal, A. and Bandyopadhyay, S. (2008). Part of Speech Tagging in Bengali Using Support Vector Machine. In *Proceedings of the 2008 International Conference on Information Technology*, pages 106–111.

Ghosh, A., Bhaskar, P., Das, A., and Bandyopadhyay, S. (2009). Dependency Parser for Bengali: the JU System at ICON 2009. In *NLP Tool Contest ICON 2009*, pages 87–91.

Hasan, K. M. A., Mondal, A., and Saha, A. (2010). A context free grammar and its predictive parser for Bangla grammar recognition. In *2010 13th International Conference on Computer and Information Technology (ICCIT)*, pages 87–91.

Hoque, M. N. and Seddiqui, M. H. (2015). Bangla Parts-of-Speech tagging using Bangla stemmer and rule based analyzer. *Proceedings of the 18th International Conference on Computer and Information Technology (ICCIT)*, pages 440–444.

Lüngen, H., Puskás, C., Bärenfänger, M., Hilbert, M., and Lobin, H., (2006). *Discourse Segmentation of German Written Texts*, pages 245–256. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

O'Donnell, M. (2000). RSTTool 2.4 – A markup tool for Rhetorical Structure Theory. In *Proceedings of the International Natural Language Generation Conference*, pages 253–256, Mizpe Ramon/Israel.

Pardo, T. A. S. and Nunes, M. d. G. V. (2008). On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *RITA*, 15:43–64.

Sarkar, P. (2006). *Bangla Byakaran Prasanga*. Dey's Publishing, Kolkata, India.

Sidarenka, U., Peldszus, A., and Stede, M. (2015). Discourse Segmentation of German Texts. *Journal for Language Technology and Computational Linguistics*, 30(1):71–98.

Taboada, M. and Mann, W. C. (2006). Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4):567–588.

Tofiloski, M., Julian, B., and Taboada, M. (2009). A Syntactic and Lexical-Based Discourse Segmenter. In *47th Annual Meeting of the Association for Computational Linguistics*, pages 77–80.

van der Vliet, N., Berzlánovich, I., Bouma, G., Egg, M., and Redeker, G. (2011). Building a discourse-annotated Dutch text corpus. In *Beyond Semantics, Bochumer Linguistische Arbeitsberichte 3*, pages 157–171.

9. Language Resource References

Cao, S., Xue, N., da Cunha, I., Iruskieta, M., and Wang, C. (2017). Discourse Segmentation for Building a RST Chinese Treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 73–81.

Cardoso, P., Maziero, E., Jorge, M. L. C., Seno, E., Di Felippo, A., Rino, L., Nunes, M. d. G., and Pardo, T. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105.

Carlson, L., Marcu, D., and Okurowski, M. E. (2002). RST Discourse Treebank, ldc2002t07.

da Cunha, I., Torres-Moreno, J.-M., and Sierra, G. (2011). On the Development of the RST Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V '11, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Iruskieta, M., Aranzabe, M. J., de Ilarraza, A. D., Gonzalez-Dios, I., Lersundi, M., and de Lacalle, O. L. (2013). The RST Basque Treebank: An online search interface to check rhetorical relations. In *Proceedings of the 4th workshop RST and discourse studies*, pages 40–49.

Stede, M. (2016). Rhetorische Struktur. In Manfred Stede, editor, *Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0*. Universitätsverlag, Potsdam.

Taboada, M. (2008). SFU Review Corpus [corpus].

Toldova, S., Pisarevskaya, D., Ananyeva, M., Kobozeva, M., Nasedkin, A., Nikiforova, S., Pavlova, I., and Shelepov, A. (2017). Rhetorical relations markers in Russian RST Treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 29–33.