Developing the Bangla RST Discourse Treebank

Debopam Das and Manfred Stede
University of Potsdam
Karl-Liebknecht Strasse 24-25,
14476 Potsdam, Germany
{debdas, stede}@uni-potsdam.de

Abstract

We present a corpus development project which builds a corpus in Bangla called the Bangla RST Discourse Treebank. The corpus contains a collection of 266 Bangla text, which are annotated for coherence relations (relations between propositions, such as *Cause* or *Evidence*). The texts represent the newspaper genre, which is further divided into eight sub-genres, such as business-related news, editorial columns and sport reports. We use Rhetorical Structure Theory (Mann and Thompson, 1988) as the theoretical framework of the corpus. In particular, we develop our annotation guidelines based on the guidelines used in the Potsdam Commentary Corpus (Stede, 2016). In the initial phase of the corpus development process, we have annotated 16 texts, and also conducted an inter-annotator agreement study, evaluating the reliability of our guidelines and the reproducibility of our annotation. The corpus upon its completion could be used as a valuable resource for conducting (cross-linguistic) discourse studies for Bangla, and also for developing various NLP applications, such as text summarization, machine translation or sentiment analysis.

Keywords: corpus, discourse annotation, coherence relations, Rhetorical Structure Theory, Bangla

1. Introduction

Coherence in discourse is, to a large extent, achieved through the use of coherence relations. Coherence relations (also known as discourse relations or rhetorical relations) refer to the semantic or pragmatic relations between text segments representing propositions, such as *Contrast* or *Elaboration*. For example, consider the following text fragment[1]:

(1) [The U.S. Coast Guard closed six miles of the Houston Ship Channel, where about 150 companies have operations,] [because the thick, black smoke obscured the area.] [wsj-1309].

In Example 1, there are two text segments (marked by square brackets) which are connected to each other by a *Reason* relation in which the second segment serves as a reason for the first one.

We are interested in creating a discourse-annotated corpus in which the texts are annotated with respect to coherence relations. Corpora with relational annotation have become increasingly useful in discourse studies, and they have been developed in various languages. These corpora have also served as valuable resources for developing various computational applications, such as discourse parsing, text summarization, or argumentation mining, to name a few.

We have recently begun to develop a corpus of Bangla texts annotated for coherence relations. We use Rhetorical Structure Theory or RST (Mann and Thompson, 1988) as the theoretical framework of the corpus, and we call the corpus the Bangla RST Discourse Treebank or Bangla RST-DT. In this paper, we present the corpus development project, describing our annotation schemes and annotation procedure. In addition, we also report on an inter-annotator agreement study, by having the initial subset of the corpus annotated by a team of annotators and evaluated thereon.

The paper is organized as follows: In Section 2., we provide a brief introduction to Rhetorical Structure Theory. Section 3. gives an account of previous works on RST-based corpora in different languages. In Section 4., we describe the annotation guidelines used to build the Bangla RST-DT. In Section 5., we provide the characteristics of the corpus, describing the training of the annotators (in Section 5.2.) and annotation procedure (in Section 5.3.). Reliability of annotation including an inter-annotator agreement study is discussed in Section 6.. Finally, Section 7. summarizes the paper, and highlights a few potential future applications of the corpus.

2. Rhetorical Structure Theory

The concept of coherence relations has been extensively studied in different discourse frameworks such as Rhetorical Structure Theory or RST (Mann and Thompson, 1988), Segmented Discourse Representation Theory or SDRT (Asher and Lascarides, 2003), the Penn Discourse Treebank or PDTB framework (Prasad et al., 2008), the Cognitive approach to Coherence Relations or CCR (Sanders et al., 1992), the Unified Linguistic Discourse Model or ULDM (Polanyi et al., 2004), or Hobbs' theory (Hobbs, 1985), further expanded by Kehler (Kehler, 2002). For our purpose, we use RST, because we believe that RST provides a healthy mix of an explanatory account of certain aspects of text organization and has proven practical applicability to a wide range of text types. In addition, RST is a language-independent theory, and it has been successfully used in a number of areas in computational discourse processing, such as text generation, discourse

[1]Example source: RST Discourse Treebank (Carlson et al., 2002). The content inside the square brackets following the example refers to the file number in the corpus from which the text fragment has been taken.

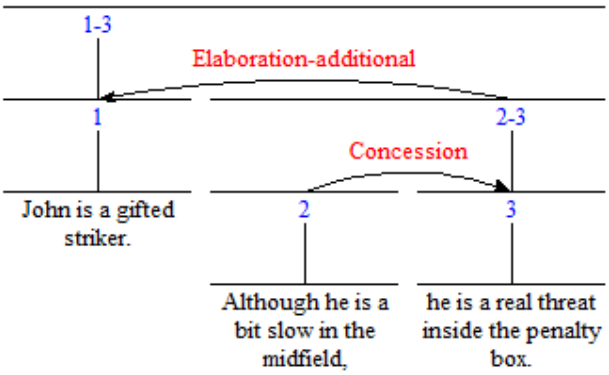

Figure 1: Graphical representation of an RST analysis

parsing, and text summarization (see Taboada and Mann (2006a) for an overview).

Rhetorical Structure Theory or RST is a functional theory of text organization (Mann and Thompson, 1988; Taboada and Mann, 2006b). It describes what parts a text is made of, what kinds of relationships exist between these parts, and how these parts are organized with respect to each other to constitute a coherent piece of discourse. In RST, relations hold between two (or sometimes more) non-overlapping text spans, and 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. The relation inventory suggested by Mann and Thompson consisted of 25 relations, but the authors emphasized that additions may be possible for specific kinds of text. Relations that are used in many projects include *Cause, Concession, Condition, Elaboration, Result* or *Summary*.

Texts, according to RST, are built out of basic clausal units (also called elementary discourse units or EDUs) that enter into rhetorical (or discourse, or coherence) relations with each other in a recursive manner. Mann and Thompson (1988) proposed that most texts can be analyzed in their entirety as recursive applications of different types of relations. In effect, this means that an entire text can be analyzed as a tree structure, with clausal units being the leavees and relations the nodes.

For illustration purposes, we provide the annotation of a short (invented) text, 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-additional* relation.

[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 tree diagrams.

3. Related work

The tradition of building discourse-annotated corpora began with the introduction of the RST Discourse Treebank or RST-DT (Carlson et al., 2002). The RST-DT contains a collection of 385 Wall Street Journal articles annotated for coherence relations. The corpus provides annotation for more than 20,000 relation instances, and the corpus has been extensively used for developing a number of RST-based discourse parsers, including Hernault et al. (2010), Ji and Eisenstein (2013), Feng and Hirst (2014) and Braud et al. (2016).

Considering the success of the RST-DT, similar corpora have been developed, in English and also in other languages. Taboada and Renkema develop the Discourse Relations Reference Corpus in English (Taboada and Renkema, 2008), annotating a set of 65 texts taken from the RST website[3], RST-DT (Carlson et al., 2002) and SFU Review Corpus (Taboada, 2008). In Dutch, an RST corpus is developed with annotation of discourse structure and also lexical cohesion (van der Vliet et al., 2011). For Brazilian Portuguese, RST is used to create the CSTNews corpus (Cardoso et al., 2011), which includes annotation of news texts and single/multi-document summaries[4]. In Spanish, da Cunha and colleagues (da Cunha et al., 2011a; da Cunha et al., 2011b) develop the RST Spanish Discourse Treebank, which includes a collection of over 250 RST-annotated texts from different specialized domains (Astrophysics, Law, Mathematics, Psychology, etc.). The Basque version of the RST corpus called the RST Basque Treebank (Iruskieta et al., 2013) is annotated not only for coherence relations, but for their signals as well. For German, the Potsdam Commentary Corpus or PCC (Stede, 2016) is built over a collection of 170 newspaper commentaries. The texts in PCC are annotated for RST relations, and also for five other different layers of annotation, such as syntax, co-reference and information structure. Initiatives to develop RST corpora have also been taken for Chinese (Cao et al., 2017) and Russian (Toldova et al., 2017), and those corpora are currently under production.

We chose to develop an RST corpus for Bangla. Bangla is an Indo-Aryan language spoken in India and Bangladesh, with an estimated 177 million speakers in the Indian subcontinent (leaving aside the diasporic Bangla speakers living elsewhere) (Dasgupta, 2003). While Bangla has remained a relatively well-studied language, unfortunately there are only a handful of linguistic corpora available for Bangla, mainly either transcribed for speech (Das et al., 2011; Bills et al., 2016), or annotated for lemmatization, POS tags or similar phenomena (Bali et al., 2010; Chaudhury et al., 2017; Ekbal and Bandyopadhyay, 2008), or unannotated (Al Mumin et al., 2014). To our knowledge, there is no discourse-annotated text corpus available

[3]http://www.sfu.ca/rst/

[4]CSTNews is also annotated based on Cross-document Structure Theory (Radev, 2000), which explains how text passages from different topics on the same topic are related to each other.

in Bangla, and in this respect, the present Bangla RST-DT is going to be the first data set of its kind.

4. Annotation Guidelines

The success of a corpus annotation task depends much on the reliability of the guidelines to be followed in the annotation. Our annotation guidelines for the Bangla RST-DT[5] are based on the guidelines used to annotate the Potsdam Commentary Corpus or PCC (Stede, 2016)[6], and are more closely related to an updated version of the PCC guidelines used in Das et al. (2017)[7]. In the present project, we adopt a modified version of these guidelines for annotating texts in Bangla, because these original guidelines (although used for German and English texts) are based on RST, which is essentially a language-independent theory.

An RST annotation of a text comprises three steps: (1) segmenting the text into EDUs (elementary discourse units), (2) assigning the relations between EDUs and larger spans, and finally (3) building the hierarchical RST tree, comprising all the connected spans stemming from a single root node at the top of the tree.

Our segmentation guidelines closely follow those used for German texts in the 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 (Mann and Thompson, 1988), which specifies that only adjunct clauses (rather than complement clauses) are considered to constitute legitimate EDUs. According to this principle, 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 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.

Since Bangla employs a different morphological and syntactic structure than German and English, it is important to determine how clause-based discourse segmentation strategies could be used for Bangla. For this purpose, we consult some notable works on Bangla grammar, such as Chatterji (1988), Chakraborty (1992), Chaki (1996) and Sarkar (2006). In addition, we closely examine the clausal structures in the Bangla texts from our corpus. Based on our understanding and observation of the Bangla syntactic structures, we ultimately decide to retain the basic segmentation principles from the PCC and SLSeg guidelines, but at the same time, we also modify some of those guidelines (or even develop some new segmentation strategies) to account for certain constructions in Bangla.

We enumerate the most significant segmentation principles followed in our annotation below. For more information about the segmentation guidelines, see (Das, under review).

1. Clausal subjects, represented as verbal nouns or complete clauses (with a finite verb) in Bangla, are not considered to be EDUs.
2. Clausal complements (including clausal objects of verbs, expressed as verbal nouns or infinitival clauses) are not considered to constitute EDUs.
3. Attribution clauses are represented either by reported speech, both directly (by direct quotes) or indirectly. They can also be represented through cognitive predicates (containing verbs expressing feelings, thoughts or opinions, such as *think*, *know*, *estimate* or *wonder* in English). Attribution clauses are not considered to form EDUs.
4. Non-restrictive relative clauses which encode a coherence relation with their host clauses are considered to be EDUs. However, restrictive relative clauses which typically elaborate on an entity in their host clauses are not considered as EDUs.
5. Participial clauses, conditional clauses, infinitival clauses (if they are not complement clauses) and verbal nouns with a postposition are treated as EDUs.

In an RST relational annotation task, the next step after segmentation is to determine the suitable coherence relations that hold between EDUs (or larger spans comprising multiple EDUs). This is done by selecting a relation type from a relational taxonomy that specifies a range of all relation types (along with their definitions) which could possibly occur in a corpus. Whenever a new relational instance in the corpus is encountered and interpreted, it is assigned a relation type which best represents the relational instance in the corpus.

The relational taxonomy used in our annotation is the one used in the PCC (Stede, 2016), which is based on the relation set proposed in the original RST paper (Mann and Thompson, 1988). This means that the relation set is much smaller than that of the RST-DT (Carlson et al., 2002) which employs a large set of 78 relations (divided into 16 broad relation classes). This is because our taxonomy does not use the many nucleus-satellite variants, and it deliberately left out suggestions like *Topic-Comment* or *Attribution*, which are not considered as coherence relations in the same way as those of "classic" RST[8].

Our taxonomy includes 31 relations which are organized in a slight different way from Mann and Thompson (1988). It

[5]The Bangla RST-DT annotation guidelines are available at: http://angcl.ling.uni-potsdam.de/pdfs/Bangla-RST-DT-Annotation-Guidelines.pdf

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

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

[8]We do not claim that phenomena of Topic-Comment and Attribution do not exist. Instead, notions of information structure in our view belong to a separate level of analysis, and not to that of coherence relations.

retains the original binary classification of subject-matter and presentational relations (semantic and pragmatic relations, respectively, in our taxonomy). We also have an extra category for *textual* relations (e.g., *List, Summary*). The taxonomy for mononuclear relations is provided in Table 1.

Semantic	Pragmatic	Textual
Circumstance	Background	Preparation
Condition	Antithesis	Restatement
Otherwise	Concession	Summary
Unless	Evidence	
Elaboration	Reason	
E-elaboration	Reason-N	
Interpretation	Justify	
Means	Evaluation-S	
Cause	Evaluation-N	
Result	Motivation	
Purpose	Enablement	
Solutionhood		

Table 1: Mononuclear relations in Bangla RST-DT

In addition, there are five multinuclear relations in our taxonomy: *Sequence, Contrast, Conjunction, List* and *Joint*. Among these, *Sequence* is a semantic relation, while the remaining four relations can function as semantic, pragmatic or textual, depending on context.

5. Corpus Development Process

5.1. Characteristics of the Corpus

The Bangla RST-DT is being developed as a corpus of Bangla 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 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 distribution of the texts for these sub-genres in the corpus is provided in Table 2.

Sub-genre	Number of texts
Business	31
Editorial column	32
Internatinal affairs	31
Cityscape	32
Letters to the editor	41
Nature	31
Science	34
Sports	34
TOTAL	266

Table 2: Distribution of texts in Bangla RST-DT

5.2. Annotator Profile and Training

The initial subset of the corpus (also used for the inter-annotator agreement study) was annotated by a team of three annotators who are native Bangla speakers. The team includes two graduate students and one of the authors of the present paper. The subset comprises 16 texts which were separately annotated by each annotator.

The two graduate students hired as the annotators have prior experience in other types of text annotation. They were extensively trained in RST by the third annotator (who has many years of experience with various RST annotation projects, and served as the *expert* annotator in this project). The training roughly consists of three phases. In the first phase, the student annotators were introduced to RST, and they learned to operate RSTTool (O'Donnell, 2000)[9] which was used for annotation. In this phase, they also did a fair amount of practice by independently annotating many texts (in English) from different genres (newspaper reports, scientific articles, undergraduate textbooks, etc.). In the second phase, the annotators were introduced to the annotation guidelines of the present project, and following those guidelines, they annotated a number of Bangla texts (which were collected from *Anandabazar Patrika*, and are similar to the target texts) as part of their practice. In the final phase, all three annotators annotated three Bangla texts separately, and compared the annotations with each other. The results were jointly discussed and adjudicated in order to resolve disagreements in annotation, arising from issues such as assigning nuclearity or choosing a relation label. The overall procedure stretched over two months, and each student annotator spent approximately 35 hours on the training phase.

5.3. Annotation Procedure

For annotating the target 16 texts in the corpus, we used pre-segmented texts. The texts were segmented beforehand by the expert annotator, following the segmentation guidelines described in the previous section. This is because we believe that segmentation is essentially a different kind of task from other tasks in relational annotation (such as deciding on mono- vs. multinuclear relations, or choosing a certain relation label). Segmentation can be factored out and evaluated separately, which has the advantage that it reduces the effort of the annotators, and makes it much easier to quantitatively evaluate the nuclearity-assignment and relation-tagging decisions (and also to perform a qualitative analysis of disagreement).

The annotation procedure consists of the following steps:

1. Identifying the macro structure: We read a text first, and identify the main topic(s), and henceforth, the most important nucleus (or nuclei). This helps us divide the text into larger units.

[9] We use the source version of RSTTool which works on Unicode scripts (with UTF-8 encoding for the Bangla script).

2. Identifying the nucleus-EDUs: We select the EDUs that play an important role in the text. If one EDU can be singled out as representing the central statement of the text, we mark it as such, and also the other important EDUs.

3. Connecting EDUs with relations: We consider each EDU and its direct neighbours, in order to see if there is 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 discourse marker. If we find such an EDU-pair, we link the EDUs with an appropriate relation, based on the nuclearity decision made in the previous step.

4. Connecting larger spans: When the EDUs are connected, they make larger spans. We link these larger spans with appropriate relations. Sometimes, the relation between larger units (and also between smaller units/EDUs) are signalled by a discourse marker, but in many cases, they can also be indicated by other textual signals (such as syntactic or lexical features) as used in the RST Signalling Corpus (Das et al., 2015). We continue to connect even larger spans until we notice all the spans are connected and we have a single root node in the tree.

In practice, as mentioned earlier, we use RSTTool to do our annotation. The texts (in .txt format) were first imported to RSTTool, and segmented by the expert annotator beforehand. The segmented texts were then distributed among all the three annotators who further did the nuclearity and relational annotation. The annotated texts are saved in .rs3 (an XML) format.

We observed that the annotators took approximately between 30 minutes and one hour to annotate a single text. The time varies mainly according to the length of the text, and also its type. For example, argumentative texts (editorial columns, letters to the editor, etc.) usually take longer time to annotate than simple news reports (business related news, sports reports, etc.).

6. Inter-annotator Agreement

In order to check the validity of our annotation guidelines and test the reproducibility of our annotation, we conducted an inter-annotator agreement study. We followed the method proposed in Marcu (2000) which evaluates agreement between competing analyses with respect to four individual dimensions: unit, span, nuclearity and relation. Since in our study we use pre-segmented texts, we calculated agreement only for span, nuclearity and relation.

We computed the agreement between a pair of annotators in terms of precision and recall. First, we calculated the agreement between the expert annotation and student annotation, considering the former as the "gold annotation". For this purpose, we made use of RSTEval, a tool that provides precision and recall statistics between a "gold" human annotation and a parser-produced annotation[10]. The results of the pairwise comparisons are provided in Table 3.

	Exp vs. Student1		**Exp vs. Student2**	
Dimension	**Precision**	**Recall**	**Precision**	**Recall**
Span	0.87	0.87	0.86	0.86
Nuclearity	0.69	0.69	0.68	0.68
Relation	0.51	0.51	0.49	0.49

Table 3: Precision and recall for expert versus student annotators

Table 3 show that the agreement between annotators is high for span, fair for nuclearity and moderate for relation This is in line with earlier studies on relational annotation (Carlson et al., 2003; Cardoso et al., 2011; da Cunha et al., 2011a), which suggest that spans are easier to identify than nuclearity status, while relation assignment is particularly difficult.

We would like to point out that the precision and recall values in a pairwise comparison are same. The identical values stem from the use of pre-segmented text files by all annotators. The same set of EDUs for a text generated the identical number of spans and also the identical number of relations across annotations. This further resulted in producing an equal number of relevant items (in the gold annotation) and retrieved items (in the annotation to be tested), which were used as the denominators for the precision and recall formula, respectively.

Second, we computed the agreement between two student annotators, again in terms of precision-recall values. Here, the annotation produced by student annotator 1 was (technically) used as the "gold annotation". The results are provided in Table 4.

Dimension	**Precision**	**Recall**
Span	0.90	0.90
Nuclearity	0.74	0.74
Relation	0.59	0.59

Table 4: Precision and recall for student annotator 1 versus student annotator 2

A comparison of all pairwise results (in Table 3 and 4) shows that the agreement is higher when the annotations produced by the student annotators were examined. This suggests that the annotators were successfully trained and were able to adhere to the annotation guidelines, which in turn yielded higher agreement between them. We thus feel that our annotations are reliable, and believe that we can use the guidelines and infrastructure to train further annotators, in order to complete the corpus annotation.

[10] http://www.nilc.icmc.usp.br/rsteval/

7. Conclusion

We have presented the development of the Bangla RST Discourse Treebank, and described our annotation guidelines and annotation procedure. The corpus started with the annotation of 16 texts, which were evaluated for agreement among the annotators. The currently-ongoing work includes annotation of the remaining 250 more texts, representative of different sub-genres in the newspaper genre.

The Bangla RST-DT when completed will be made publicly available. The corpus will have two clear applications. First, from a theoretical point of view, it will provide empirical support to the existing research on the discourse structure of Bangla, an Indo-Aryan language, and provide a valuable resource for conducting cross-linguistic discourse studies. Second, the corpus will be used to develop discourse parsing systems for Bangla texts, which may further be used for NLP applications such as automatic summarization, machine translation, sentiment analysis, or argumentation mining.

Furthermore, since our annotation guidelines have now been successfully applied to German, English, and Bangla, with only minor changes, we believe that they are now quite stable across languages and can serve for projects by other researchers, possibly involving further languages.

8. Acknowledgements

We would like to thank our two student annotators, Lahari Chatterjee and Soumya Sankar Ghosh, for their valuable contribution to the development the Bangla RST Discourse Treebank.

9. Bibliographical References

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press, Cambridge.

Braud, C., Plank, B., and Søgaard, A. (2016). Multi-view and multi-task training of RST discourse parsers. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 1903–1913, Osaka, Japan.

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.

Dasgupta, P. (2003). Bangla. In George Cardona et al., editors, *The Indo-Aryan Languages*, pages 386–427. Routledge.

Das, D. (under review). Discourse Segmentation in Bangla. In *4th Workshop on Indian Language Data: Resources and Evaluation (WILDRE-4)*.

Feng, V. W. and Hirst, G. (2014). A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 511–521, Baltimore, MA.

Hernault, H., Prendinger, H., duVerle, D., and Ishizuka, M. (2010). Hilda: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3):1–33.

Hobbs, J. (1985). On the coherence and structure of discourse. Report, CSLI.

Ji, Y. and Eisenstein, J. (2013). Discriminative improvements to distributional sentence similarity. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 891–896, Seattle, Washington, USA, October. Association for Computational Linguistics.

Kehler, A. (2002). *Coherence, Reference, and the Theory of Grammar*. Center for the Study of Language and Information, Stanford, CA.

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

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

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.

Polanyi, L., Culy, C., van den Berg, M., Thione, G. L., and Ahn, D. (2004). A rule based approach to discourse parsing. In *the 5th SIGdial Workshop on Discourse and Dialogue, ACL*.

Radev, D. R. (2000). A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-document Structure. In *Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue - Volume 10*, SIGDIAL '00, pages 74–83, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sanders, T., Spooren, W., and Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15:1–35.

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

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

Taboada, M. and Mann, W. C. (2006b). Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

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.

10. Language Resource References

Al Mumin, M. A., Shoeb, A. A. M., Selim3, M. R., and Iqbal, M. Z. (2014). Sumono: A representative modern

bengali corpus. *SUST Journal of Science and Technology*, 21(1):78–86.

Bali, K., Choudhury, M., and Biswas, P. (2010). Indian language part-of-speech tagset: Bengali ldc2010t16.

Bills, A., David, A., Dubinski, E., Fiscus, J., Gillies, B., Harper, M., Jarrett, A., Molina, M., Ray, J., Rytting, A., Paget, S., Shen, W., Silber, R., Tzoukermann, E., and Wong, J. (2016). Iarpa babel bengali language pack iarpa-babel103b-v0.4b.

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.

Carlson, L., Marcu, D., and Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt et al., editors, *Current Directions in Discourse and Dialogue*, pages 85–112. Kluwer, Dordrecht.

Chaudhury, T. H., Matin, A., Hossain, M., Uzzaman, A., and Masum, M. (2017). Annotated bangla news corpus and lexicon development with pos tagging and stemming. *Global Journal of Researches in Engineering*, 17(1).

da Cunha, I., Torres-Moreno, J.-M., and Sierra, G. (2011a). 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.

da Cunha, I., Torres-Moreno, J.-M., Sierra, G., Cabrera-Diego, L. A., Rolón, B. G. C., and Bartilotti, J. M. R. (2011b). The Rst Spanish Treebank On-line Interface. In *Proceedings of the Recent Advances in Natural Language Processing*, pages 698–703.

Das, B., Mandal, S., and Mitra, P. (2011). Bengali speech corpus for continuous automatic speech recognition system. In *Proceedings of Conf. Speech Database and Assessments (Oriental COCOSDA)*, pages 51–55.

Das, Debopam and Taboada, Maite and McFetridge, Paul. (2015). *RST Signalling Corpus, LDC2015T10*. Linguistic Data Consortium.

Ekbal, A. and Bandyopadhyay, S. (2008). Web-based bengali news corpus for lexicon development and pos tagging. *Polibits*, 37.

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.

Prasad, Rashmi and Dinesh, N. and Lee, A. and Miltsakaki, E. and Robaldo, L. and Joshi, A. and Webber, B. (2008). *The penn discourse treebank 2.0*.

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

Taboada, M. and Renkema, J. (2008). Discourse relations reference corpus.

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.

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.