Discourse Relations Reference Corpus

Maite Taboada

Department of Linguistics
Simon Fraser University
8888 University Dr.
Burnaby
B.C. V5A 1S6
Canada

mtaboada@sfu.ca

Jan Renkema

Tilburg University
Faculty of Arts
Dantebuilding room 420
PO Box 90153
5000 LE Tilburg
The Netherlands

j.renkema@uvt.nl

1 Introduction

The Discourse Relations Reference Corpus is a resource for those interested in the study of discourse relations from a corpus linguistics perspective. By discourse relations we mean rhetorical or coherence relations that hold among pieces of discourse, and that create the impression of coherence. Different approaches to discourse relations are discussed in Mann and Thompson (1988), Polanyi (1988), Sanders et al. (1993), Renkema (2004), Asher and Lascarides (2003), Taboada and Mann (2006), and references therein.

The materials in this corpus are taken from three different sources: texts from the RST web site (Mann and Taboada, 2007); annotated Wall Street Journal articles from the RST Discourse Treebank (Carlson et al., 2002); and review texts from the SFU Review Corpus (Taboada, 2008). We provide information about each one of these in the following sections, and further links to other corpora relevant to discourse relations research.

The documents in each of the subcorpora have been annotated with RSTTool (http://www.wagsoft.com/RSTTool), a program that provides a graphic interface to annotate relations, indicating nucleus and satellite status and providing a label, which can be chosen from the standard RST set, or from a purpose-built one.

Although the background to all subcorpora is Rhetorical Structure Theory, and they have been annotated with RSTTool, we believe that the corpus is useful to anyone interested in discourse relations, from whatever perspective. The annotations provide rich information on what relations are more common; how they are commonly signalled; and how relations are distributed in different genres.

2 RST website corpus

The RST website (http://www.sfu.ca/rst) contains information about Rhetorical Structure Theory, including definitions for all relations, bibliographies, and links to tools and other resources. The site also contains examples of RST analyses, divided into published and

unpublished analyses. Published texts have appeared in print in the publications about RST by its creators (Mann et al., 1992; Mann and Thompson, 1983, 1986, 1987, 1988, 1992; Thompson and Mann, 1987a, 1987b). The unpublished analyses were also collected and analyzed by the creators of RST.

We have included in the Discourse Relations Reference Corpus all of the texts from the RST site. They are 15 texts, one of them analyzed in three different ways.

3 RST Discourse Treebank corpus

The RST Discourse Treebank corpus (Carlson et al., 2002) is a collection of Wall Street Journal articles annotated according to a version of RST. The corpus contains 385 articles published in the Wall Street Journal, a subset of the large Penn Treebank (Marcus et al., 1999). A portion of the articles was also annotated following a graph structure (Wolf et al., 2005). This corpus was annotated with a large relation set, 78 relations in total. The entire corpus is available through the Linguistic Data Consortium (http://www.ldc.upenn.edu/).

We included a subset of 30 texts in our reference corpus. They were selected because an abstract was also included with the LDC corpus release. (We do not include the abstracts in our corpus; only the full texts). The abstracts were used to evaluate a summarization system based on the RST analyses.

The texts were annotated with a version of the RSTTool: http://www.isi.edu/licensed-sw/RSTTool/. More information on the corpus annotation and reliability measures can be found in publications by Marcu, Carlson and colleagues (Carlson and Marcu, 2001; Carlson et al., 2001, 2003).

4 SFU Review Corpus

As part of a project on extracting sentiment from text, a team of researchers at Simon Fraser University, led by Maite Taboada, collected a corpus of movie, book, and consumer product reviews. More information on the corpus collection, and the project it is part of, can be obtained from: http://www.sfu.ca/~mtaboada/research/nserc-project.html, and from several publications (Taboada et al., 2006; Taboada and Grieve, 2004). The reviews were downloaded in 2004 from the Epinions web site (http://www.epinions.com). They are divided into eight categories, with 25 positive and 25 negative reviews in each category, i.e., reviews that contained an overall positive or negative evaluation of the product. The classification into positive and negative was based on the "recommended" or "not recommended" tag that the reviewer provided. The categories are: books, cars, computers, cookware, hotels, movies, music, and phones.

A subset of the corpus is included in the Discourse Relations Reference Corpus, 15 reviews in the movie and book categories, for a total of 30 texts. The annotations are subsentential, i.e., only relations within the sentence are annotated. The corpus was annotated by only one annotator (Montana Hay), with input from Maite Taboada. This corpus has not been checked for consistency or accuracy, and it should not be considered

a standard. We provide it only as another source of relation examples in a specific genre. If the annotations are checked and compared in the future, we will release a new version.

More information on the corpus, including texts in all categories, the RST annotations, and a small set of Appraisal (Martin and White, 2005) annotations, is available on the web: http://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html.

5 Analysis files

This directory contains the analysis files, which can be opened using RSTTool. Note that the Marcu RST Treebank files have a .lisp extension, but will still open with RSTTool.

6 Other resources

The RST web site contains further information on Rhetorical Structure Theory and a bibliography of related materials: http://www.sfu.ca/rst/.

For German, Stede and colleagues have created a corpus of German newspaper editorials annotated according to RST (Stede, 2004).

7 Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada, under a Discovery Grant to Maite Taboada. We would like to thank Frank Dortmans, MA student at Tilburg University, and Simon Fraser University students Jack Grieve, Montana Hay and Patrick Larrivee-Woods for their contribution to the corpus compilation and annotation.

References

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

Carlson, Lynn and Daniel Marcu. (2001). *Discourse Tagging Manual*. Unpublished manuscript, http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf.

Carlson, Lynn, Daniel Marcu and Mary Ellen Okurowski. (2001). Building a discourse tagged corpus in the framework of Rhetorical Structure Theory. *Proceedings of 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*. Aalborg, Denmark.

Carlson, Lynn, Daniel Marcu and Mary Ellen Okurowski. (2002). RST Discourse Treebank, LDC2002T07 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Carlson, Lynn, Daniel Marcu and Mary Ellen Okurowski. (2003). Building a discourse tagged corpus in the framework of Rhetorical Structure Theory. In J. van

Kuppevelt and R. Smith (Eds.), *Current and New Directions in Discourse and Dialogue* (pp. 85-112). Berlin: Springer.

Mann, William C., Christian M.I.M. Matthiessen and Sandra A. Thompson. (1992). Rhetorical Structure Theory and text analysis. In W. C. Mann and S. A. Thompson (Eds.), *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text* (pp. 39-78). Amsterdam and Philadelphia: John Benjamins.

Mann, William C. and Maite Taboada. (2007). *RST Web Site*. Retrieved July 2008, from http://www.sfu.ca/rst

Mann, William C. and Sandra A. Thompson. (1983). *Relational Propositions in Discourse* (Technical Report No. ISI/RR-83-115). Marina del Rey, CA: Information Sciences Institute.

Mann, William C. and Sandra A. Thompson. (1986). *Rhetorical Structure Theory: Description and Construction of Text Sructures* (Technical Report No. ISI/RS-86-174). Marina del Rey, CA: Information Sciences Institute.

Mann, William C. and Sandra A. Thompson. (1987). *Rhetorical Structure Theory: A Theory of Text Organization* (No. ISI/RS-87-190). Marina del Rey, CA: Information Sciences Institute.

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

Mann, William C. and Sandra A. Thompson. (1992). Relational Discourse Structure: A comparison of approaches to structuring text by 'contrast'. In S. J. J. Hwang and W. R. Merrifield (Eds.), *Language in Context: Essays for Robert E. Longacre* (pp. 19-45). Dallas: Summer Institute of Linguistics and the University of Texas at Arlington.

Marcus, Mitchell P., Beatrice Santorini, Mary Ann Marcinkiewicz and Ann Taylor. (1999). Treebank-3, LDC99T42 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Martin, James R. and Peter White. (2005). *The Language of Evaluation*. New York: Palgrave.

Polanyi, Livia. (1988). A formal model of the structure of discourse. *Journal of Pragmatics, 12*, 601-638.

Renkema, Jan. (2004). *Introduction to Discourse Studies*. Amsterdam and Philadelphia: John Benjamins.

Sanders, Ted, Wilbert Spooren and Leo Noordman. (1993). Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics, 4* (2), 93-133.

Stede, Manfred. (2004). The Potsdam commentary corpus. *Proceedings of the Workshop on Discourse Annotation, 42nd Meeting of the Association for Computational Linguistics*. Barcelona, Spain.

Taboada, Maite. (2008). SFU Review Corpus [Corpus]. Vancouver: Simon Fraser University, http://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html.

Taboada, Maite, Caroline Anthony and Kimberly Voll. (2006). Methods for creating semantic orientation dictionaries. *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)* (pp. 427-432). Genoa, Italy.

Taboada, Maite and Jack Grieve. (2004). Analyzing appraisal automatically. *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS-04-07)* (pp. 158-161). Stanford University, CA.

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

Thompson, Sandra A. and William C. Mann. (1987a). Antithesis: A study in clause combining and discourse structure. In R. Steele and T. Threadgold (Eds.), *Language Topics: Essays in Honour of Michael Halliday, Volume II* (pp. 359-381). Amsterdam and Philadelphia: John Benjamins.

Thompson, Sandra A. and William C. Mann. (1987b). Rhetorical Structure Theory: A framework for the analysis of texts. *IPrA Papers in Pragmatics, 1* (1), 79-105.

Wolf, Florian, Edward Gibson, Amy Fisher and Meredith Knight. (2005). Discourse GraphBank, LDC2005T08 [Corpus]. Philadelphia: Linguistic Data Consortium.