Multi-Layer Discourse Annotation of a Dutch Text Corpus

Gisela Redeker,[*] Ildikó Berzlánovich,[*] Nynke van der Vliet,[*] Gosse Bouma,[*] Markus Egg[†]
[*] University of Groningen, † Humboldt University, Berlin
Groningen, The Netherlands; Berlin, Germany
E-mail: {g.redeker,i.berzlanovich,n.h.van.der.vliet,g.bouma}@rug.nl, markus.egg@anglistik.hu-berlin.de

Abstract

We have compiled a corpus of 80 Dutch texts from expository and persuasive genres, which we annotated for rhetorical and genre-specific discourse structure, and lexical cohesion with the goal of creating a gold standard for further research. The annotations are based on a segmentation of the text in elementary discourse units that takes into account cues from syntax and punctuation.

During the labor-intensive discourse-structure annotation (RST analysis), we took great care to thoroughly reconcile the initial analyses. That process and the availability of two independent initial analyses for each text allows us to analyze our disagreements and to assess the confusability of RST relations, and thereby improve the annotation guidelines and gather evidence for the classification of these relations into larger groups.

We are using this resource for corpus-based studies of discourse relations, discourse markers, cohesion, and genre differences, e.g., the question of how discourse structure and lexical cohesion interact for different genres in the overall organization of texts. We are also exploring automatic text segmentation and semi-automatic discourse annotation.

Keywords: discourse structure, coherence relations, lexical cohesion

1. Introduction

Texts are structured entities that exhibit coherence and cohesion. Much research on coherence targets local coherence relations and their linguistic signaling (Sanders et al., 1992; Knott & Sanders, 1998; Prasad et al., 2008). Configurational issues concerning the hierarchical structure of texts, e.g., their complexity (Stede, 2004; Mann & Thompson, 1988; Wolf & Gibson, 2005), are widely discussed, but still lack a substantial empirical foundation. The interplay of relational discourse structure with cohesion was investigated with a focus on anaphora interpretation (Fox, 1987; Poesio et al., 2004), while the role of lexical cohesion in the overall textual organization received little attention (except Hasan, 1984; Hoey, 1991).

Textual organization depends on genre (Eggins & Martin, 1997; Webber, 2009). In particular, persuasive texts are organized around a central purpose or intention, while descriptive or expository texts are usually organized around a theme, moving through sub-themes. Since this difference affects relational structure and lexical cohesion, our corpus covers different genres. The genre-specific structure of a text can be described by its *moves* (genre-specific main functional text components; Biber et al. 2007). Conventionalized genres have a prototypical or canonical (though not completely rigid) move pattern.

By annotating relational and lexical organization in a variety of genres, we have created a Dutch language resource for corpus-based discourse research, computational modeling, and applications like summarization.

2. Corpus Design

Our aim is to create a reliable "gold standard" resource covering genres from two classes: expository texts, which present information, and persuasive texts, which aim to affect readers' intentions or actions. The expository subcorpus comprises 20 entries from two online encyclopedias[1] (labeled EE in the corpus) and 20 from a science news website[2] (PSN). The persuasive texts are 20 fundraising letters (FL) and 20 commercial advertisements from magazines (AD). The texts have 190-400 words.

The annotation for discourse structure, moves, and lexical cohesion (see sections 3 and 4) is based on a segmentation of the texts into elementary discourse units (EDUs) similar to Tofiloski et al. (2009). The segmentation rules use syntax and punctuation (van der Vliet et al., 2011) and were implemented in an automatic segmenter (van der Vliet, 2010). We used O'Donnell's (1997) RSTTool for the annotation of discourse structure and an MMAX-based tool (Müller & Strube, 2001) for the annotation for lexical cohesion. The interplay between the various annotation layers is discussed in more detail in section 6.

All annotations were done separately by two annotators and then reconciled, guaranteeing a high degree of intersubjectivity. For the separate analyses, inter-annotator agreement for 16 texts (four of the 20 per genre) showed Kappa values that represent substantial to almost perfect agreement according to the scale of Landis and Koch (1977).

The agreement on the identification of segment boundaries was 0.97. For the RST analysis, the agreement on discourse spans was 0.83, the agreement on the labeling of nuclearity 0.77 and the agreement on the labeling of RST relations 0.70. For the move analysis, the agreement on the identification of move boundaries was 0.76 and the agreement on the labeling of the moves 0.87. For the lexical cohesion analysis, the agreement on the identification of relations between items is 0.86 and the agreement on the labeling of these relations 0.91.

[1] http://www.astronomie.nl; http://www.sterrenwacht-mercurius. nl/encyclopedie.php5
[2] http://www.scientias.nl/category/astronomie

3. Discourse Structure

The annotation of discourse structure targets the hierarchical structures arising from the recursive application of coherence relations between discourse units, and genre-specific structures crucial for understanding genre differences in discourse structure.

3.1 RST Analysis

We analyze coherence structures with Rhetorical Structure Theory (RST; Mann & Thompson, 1988; Taboada & Mann, 2006a). [3] RST has proven successful for the analysis of texts in various languages (see Taboada & Mann, 2006a,b) and the annotation of large text corpora (Carlson et al., 2002; Stede, 2004).

The use of coherence relations differs significantly between the four genres in our corpus. In a discriminant analysis, eight relations proved good predictors of genre, correctly classifying 69 of the 80 texts (86.3%) in a cross-validated analysis. Two discriminant functions (linear combinations of the variables optimized to explain between-group variance) have eigenvalues above 1. Figure 1 shows that the first discriminant function distinguishes expository (EE, PSN) from persuasive (FL, AD) genres; the second marks the difference between the mainly descriptive encyclopedia texts (EE) and the more explanatory popular science news texts (PSN).

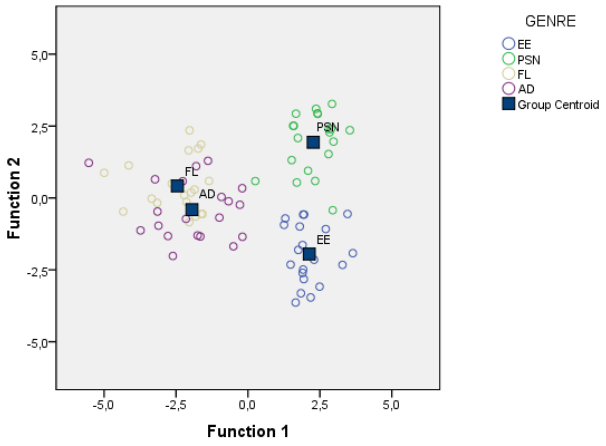

Figure 1: Clustering of texts (discriminant analysis)

Discourse-annotated corpora are particularly useful for investigating the realizations, linguistic marking, and genre-specific uses of coherence relations (e.g., Webber, 2009; Taboada et al., 2009) and we are researching such questions with our corpus. Since we also investigate the configurational characteristics of discourse structure, we represent the full hierarchical structure of our texts.

3.1.1 Confusability of RST Relations

In ongoing work, we are using the initial RST analyses to investigate the confusability of RST relations between annotators and in relation to the final, reconciled annotation.

[3] See http://www.sfu.ca/rst/ for definitions of the RST relations.

Detailed analyses of the disagreements will be used to refine our coding manual with supplementary instructions and atypical examples. For instance, most hypotactic RST relations show a preferred order of nucleus and satellite. Annotator agreement tends to be lower for relations in non-preferred order, presumably reflecting a base-rate bias. For some relations, however, there are subtle meaning differences in the non-preferred order. A post-posed Concession satellite, for instance, suggests an afterthought if the satellite occurs in a new sentence.

The Elaboration relation in Figure 2 illustrates the confusability of Elaboration with Circumstance (annotator1) and Background (annotator2). The confusion with Circumstance only occurred with Elaboration relations in the non-preferred satellite-nucleus order.

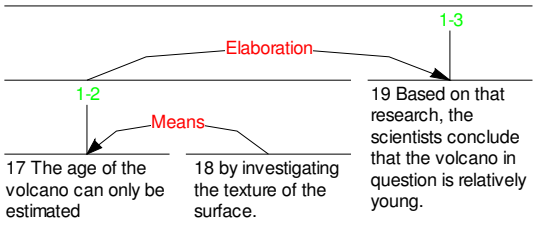

Figure 2: Elaboration relation (non-canonical order)

As far as we know, this is the first attempt to systematize and refine annotation guidelines through the systematic analysis of annotator disagreement.

Another aim of the confusability analysis is the assessment of proposals for the ordering of the relations into broader types or categories (Mann & Thompson, 1988; Carlson & Marcu, 2001; Prasad et al., 2008) or in a taxonomic system (Sanders et al., 1992). We interpret the confusability of two relations as a measure of their similarity, which is then to be spelled out in terms of common feature values or classes of relations.

Merging previous proposals, we tentatively propose the following classification (Table 1):

Expansion Relations	Semantic Relations	Pragmatic Relations
background	condition	antithesis
circumstance	means	concession
elaboration	non-volitional cause	enablement
evaluation	non-volitional result	evidence
interpretation	otherwise	justify
preparation	purpose	motivation
restatement	solutionhood	
summary	unconditional	
conjunction*	unless	
disjunction*	volitional.cause	
joint*	volitional.result	
list*	contrast*	
restatement-mn*		
sequence*		

* *multinuclear relations*

Table 1: Relation Types

Our results so far suggest that (i) confusability is genre dependent and (ii) pragmatic (presentational, intentional) relations are not usually confused with semantic or expansion relations. This is in line with Sanders (1997), who found substantial agreement in classifying relations as semantic or pragmatic and strong contextual (genre) effects for less agreed-on instances.

3.2 Genre Analysis

To compare the global text structure across genres, we combine genre analysis with RST (Taboada & Lavid, 2003; Gruber & Muntigl, 2005). We identified the genre-specific moves and overlayed the RST-tree with a segmentation into a sequence of moves. Moves partition the EDUs in the text and are realized by at least one complete EDU (contrary to Biber et al., 2007). The move types in encyclopedia entries are *name*, *define* and *describe*; those for science news texts were adapted from Haupt (2010). For fundraising letters, we followed Upton (2002), for advertisements we adapted Bhatia (2005). Figure 3 illustrates the mapping of the moves onto the RST tree for one of the fundraising letters.

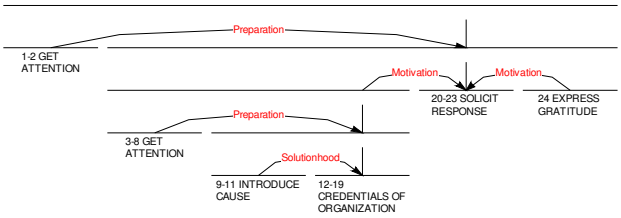

Figure 3: Move analysis mapped onto RST structure

4. Lexical Cohesion

Our analysis of lexical cohesion (Halliday & Hasan, 1976; Tanskanen, 2006) classifies the semantic relations among lexical items in the text as *repetition* (fully or partial), *systematic semantic relation* (like hyponymy, meronymy, or antonymy), or *collocation*. Items participating in lexical cohesion include content words (nouns, verbs, adjectives, and adverbs of place, time, and frequency) and proper names.

Consider the following example from one of the encyclopedia texts:

EDU5[After the **forming** of the **sun** and the **solar system**, our **star** began its long existence as a so-called **dwarf star**.] EDU6[In the **dwarf phase** of its **life**, the energy that the **sun** gives off is generated in its core through the fusion of hydrogen into helium.] EDU7[The **sun** is about five billion years old now.]

Figure 4: Lexical cohesion relations

Note that we include only relations across, not within, EDUs. In the above example, this means that the lexical relations between *sun*, *solar system, star,* and *dwarf star* in EDU5 are not included in our analysis. This allows us to investigate the co-occurrence of lexical cohesion types with coherence relations, and the alignment between discourse structure and lexical cohesion, as both structures are based on the same units.

5. Corpus-based studies

The rich annotation of our corpus allows us to investigate the interplay of rhetorical and genre-specific discourse structure, lexical cohesion, and discourse markers. We find the expected genre differences in the use of coherence relations, with pragmatic relations abounding in persuasive texts, and almost absent from expository texts, and significantly more systematic semantic lexical cohesion relations in expository than in persuasive texts (Berzlánovich & Redeker, 2011, 2012; Berzlánovich, Egg, & Redeker, in press).

We tested the hypothesis that cohesion contributes differently to textual organization in different genres: substantially in expository texts (Morris & Hirst, 1991), but minimally in persuasive texts. If lexical cohesion cues coherence structure, a high density of lexical cohesion relations, indicating centrality of the discourse unit they are associated with, should be correlated with centrality in the hierarchical coherence structure (indicated by a high level in the RST-tree). Figure 5 shows the mean lexical densities for moves at various levels in the RST tree for the four genres (using the reciprocal of the depth of embedding as a centrality score).

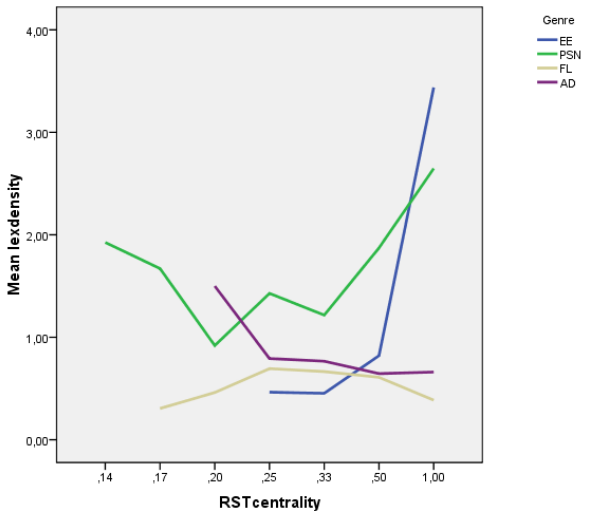

Figure 5: Coherence and Lexical Cohesion

In the expository texts (EE and PSN), the correlation between RST centrality and lexical density of the moves is .59 ($p < .001$), in the persuasive texts (FL and AD) it is -.12 ($p = .019$) (Berzlánovich & Redeker, 2011).

Coherence relations and genre-specific moves can be marked by lexical or phrasal *discourse markers*. Some relations are often marked, others seldom (Taboada 2006). Van der Vliet and Redeker (2011) analyze the discourse marker use in our corpus. The most striking result is the difference in the extent of explicit marking within (69%) and between (16%) sentences. Closer analyses will investigate differences between relation types and the extent to which the explicit marking of intra-sentential relations reflects syntactic requirements to combine clauses by conjunctions or adverbs.

6. Managing multi-layer annotation

All our annotations are available as XML, but as the various layers have been created by different tools using both in line and stand-off annotation, the XML is difficult to use and explore in combination. Some of the issues that arose during construction of the corpus are: ensuring consistency of character encoding, spelling, and tokenization, adequate representation of word order (the discourse annotation tool does not allow annotation of 'embedded' discourse segments in a way that respects the original word order), and appropriate XML encoding of various 'auxiliary' levels of annotation (discourse segmentation, discourse moves, and document lay-out). In addition, all annotation is in proprietary formats. As a consequence, it is difficult to understand the organization of the raw data, the significance of certain elements and attributes used in the XML, and especially, how various annotation layers are connected.

In this section, we describe in more detail how various annotation layers are connected, and our plans for converting the present heterogeneous annotation into a single XML format.

Text has been normalized to UTF-8, tokenized and segmented into sentences using the Alpino tools.[4] The RST annotation is created using O'Donnell's RST tool.[5] The MMAX annotation tool[6] was used to mark pairs of lexical items as expressing a lexical cohesion relation. The output of these annotation tools is always XML, but alignment and integration of the various annotation layers is non-trivial. Conceptually, the RST discourse relations form a tree over the input text. A complication with our RST trees is that they do not always follow the original word order:

(1) Op deze manier heeft Kepler - die begin 2009 werd gelanceerd - nu al vijf exoplaneten ontdekt.
In this way, Kepler - which was launched in early 2009 - has by now already discovered five exoplanets

In this example, the relative clause is annotated as an EDU which is a dependent of the EDU formed by the rest of the sentence. Such 'embedded' EDUs are not properly supported by the RSTTool. A solution is to place the embedded EDU after the main clause, and to insert a placeholder indicating the original position of the removed EDU. This *ad-hoc* solution does allow annotators to complete the annotation according to their linguistic principles, but causes serious problems when combining the annotation with other layers.

The 'in-line' annotation of both the RST-tool and Alpino XML also makes it hard to combine annotations. Although EDUs tend to be clausal in nature, it does not mean that EDUs align easily with syntactic constituents. In the example above, for instance, syntax considers *Kepler - die begin 2009 werd gelanceerd* as a constituent, but in the RST the relative clause forms an EDU, while the name *Kepler* is part of the main EDU. This suggests that an XML format combining the annotations should use some form of 'stand-off' annotation, where tokens are the base data, and pointers are used to connect the linguistic annotation to the base data.

Lexical cohesion relations, finally, establish directed links between sequences of tokens similar to coreference chains. The MMAX annotation tool that was used for lexical cohesion was designed for multi-layer annotation, and has the advantage that it provides stand-off annotation. Sequences of tokens (that can be discontinuous, in contrast with the RSTTool) are annotated as markables. Markables can be linked to each other, with labels expressing the nature of the relation. Conceptually, it is straightforward to convert the RST annotation to the more general MMAX format. Initial experiments with such a conversion have already been carried out.

In the near future, we plan to convert all annotation layers into a single XML format that properly separates base data (tokens) from higher annotation layers. In the linguistic annotation, we can distinguish between layers that segment the input (into sentences, paragraphs, EDUs, and discourse moves) and layers that add relations between segments (RST discourse relations, lexical relations, and syntactic dependency relations). Segmentation basically requires defining spans over the base data, while higher levels of annotation can be defined as labeled links between text spans.

The first thing novel users of a corpus want to do, is browse and explore the data and its annotation. While we do not envisage the development of sophisticated search and visualization tools, we do believe that some support is desirable. The ANNIS software,[7] for instance, supports import of data in MMAX, RSTTool, and TIGER format for syntactic annotation. By converting our data into this format, we obtain sophisticated visualization options.

7. Conclusion and Outlook

Our corpus aims at a high standard of empirical validity and coverage across a theoretically motivated selection of genres. With its 80 core texts, it is large enough for distributional analyses and structural comparisons. As our coherence annotation follows the widely used "classic" RST, our corpus supports cross-linguistic research by its compatibility with RST-based corpora in other languages. We are preparing detailed manuals documenting our annotations and will integrate the various XML formats of our annotation layers to facilitate distribution and use of our corpus.

Van der Vliet is exploring the combined use of our annotation layers and a list of discourse markers for developing a semi-automatic parsing tool for coherence relations. Manual annotation of discourse markers, as in the Penn Discourse TreeBank (Prasad et al. 2008), is also considered, but would require sense disambiguation and scoping rules compatible with structures and labels in our

[4] www.let.rug.nl/vannoord/alp/Alpino

[5] www.wagsoft.com/RSTTool/

[6] mmax2.sourceforge.net/

[7] www.sfb632.uni-potsdam.de/d1/annis/

RST-trees.

We envisage combining our lexical cohesion analysis with computational coreference resolution (Hendrickx et al. 2008). This will allow us to test our network model of lexical cohesion against lexical chaining approaches (e.g., Barzilay & Elhadad 1997), enhancing the value of our corpus for empirical and theoretical work. Our twofold approach to centrality (in coherence and cohesion) makes our corpus a valuable resource for applications like summarization or sentiment analysis: Centrality can, for instance, provide scores for summary-worthiness (Marcu 2000) or weigh evaluative expressions (Voll and Taboada 2007).

8. Acknowledgements

The work reported here is supported by grant 360-70-280 of the Netherlands Organization for Scientific Research (NWO). Online documentation of the program *Modeling textual organization: Discourse structure and cohesion* is available from www.let.rug.nl/mto.

9. References

Barzilay, R., Elhadad, M. (1997). Using lexical chains for text summarization. In *Proceedings of the ACL'97/ EACL'97 workshop on intelligent scalable text summarization*. Madrid, pp.10-17.

Berzlánovich, I., Egg, M., Redeker, G. (in press). Coherence structure and lexical cohesion in expository and persuasive texts. In A. Benz, P. Kühnlein, M. Stede (Eds.), *Constraints in Discourse 3*. Benjamins New Series on Pragmatics and Beyond. Amsterdam: Benjamins.

Berzlánovich, I., Redeker, G. (2011). A corpus-based investigation of coherence and lexical cohesion. *12th International Pragmatics Conference*. Manchester, 2011.

Berzlánovich, I., Redeker, G. (2012). Genre-dependent interaction of coherence and lexical cohesion in written discourse. *Corpus Linguistics and Linguistic Theory*. 8(1), pp. 183-208.

Biber, D., Connor, U., Upton, Th. (2007). *Discourse on the move*. Amsterdam: Benjamins.

Carlson, L. Marcu, D. (2001). Discourse tagging reference manual. *ISI Technical Report ISI-TR 545*.

Carlson, L., Marcu, D., Okurowski, M. (2002). *RST Discourse Treebank*. Philadelphia, PA: Linguistic Data Consortium.

Eggins S., Martin, J. (1997). Genres and registers of discourse. In T. van Dijk (Ed.), *Discourse as Structure and Process*, London: Sage, pp. 230–257.

Fox, B. (1987). *Discourse structure and anaphora*. Cambridge: Cambridge University Press.

Gruber, H., Muntigl, P. (2005). Generic and rhetorical structures of texts: Two sides of the same coin? *Folia Linguistica*, 39, pp. 75–113.

Halliday, M.A.K., Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hasan, R. (1984). Coherence and cohesive harmony. In J. Flood (Ed.), *Understanding reading comprehension: Cognition, language and the structure of prose*. Newark, DE: International Reading Association, pp. 181–219.

Haupt, J. (2010). Palpated, phonendoscoped, x-rayed and tomographed: The structure of science news in good shape. In R. Jančaříková (Ed.), *Interpretation of Meaning Across Discourses,* Masaryk University, pp. 161–174.

Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.-M., Van der Vloet, J., Verschelde, J.-L. (2008). A coreference corpus and resolution system for Dutch. In *Proceedings of LREC 2008*.

Hoey, M. (1991). *Patterns of lexis in text.* Oxford: Oxford University Press.

Knott, A., Sanders, T. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30, pp. 135–175.

Landis, J., Koch G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, pp. 159–174.

Mann, W., Thompson, S. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8, pp. 243–281.

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. Cambridge, MA: MIT Press.

Morris, J., Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17, pp. 21–48.

Müller, Ch., Strube, M. (2001). MMAX: A tool for the annotation of multi-modal corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Seattle, Wash., 5 August 2001, pages 45-50.

O'Donnell, M. (1997). RST-Tool: An RST analysis tool. In *Proceedings of the 6th European Workshop on Natural Language Generation*, March 24–26, 1997 Duisburg, Germany: Gerhard-Mercator University.

Poesio, M., Stevenson, R., DiEugenio, B., Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30, pp. 309–363.

Prasad, R. Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L. Joshi, A. Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008),* Marrakech, Morocco.

Sanders, T. (1997). Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes,* 24, pp. 119-147.

Sanders, T., Spooren, W., Noordman, L. (1992). Towards a taxonomy of coherence relations. *Cognitive Linguistics*, 15, pp. 1–35.

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, 25–26 July 2004, Barcelona, Spain, pp. 96-102.

Taboada, M. (2006). Discourse markers as signals (or not)

of rhetorical relations. *Journal of Pragmatics*, 38, pp. 567–592.

Taboada, M., Lavid, J. (2003). Rhetorical and thematic patterns in scheduling dialogues. *Functions of Language*, 10, pp. 147–178.

Taboada, M., Mann, W. (2006a). Applications of rhetorical structure theory. *Discourse Studies*, 8, pp. 567–588.

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

Taboada, M., Brooke, J., Stede, M. (2009). Genre-based paragraph classification for sentiment analysis. In *Proceedings of 10th Annual SIGDIAL Conference on Discourse and Dialogue*. London, UK. September 2009. pp. 62-70.

Tanskanen, S.-A. (2006). *Collaborating towards coherence: Lexical cohesion in English discourse*. Amsterdam: Benjamins.

Tofiloski, M., Brooke, J., Taboada, M. (2009). A syntactic and lexical-based discourse segmenter. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*. Singapore, August 2009. pp. 77-80.

Upton, Th. (2002). Understanding direct mail letters as a genre. *International Journal of Corpus Linguistics*, 7, pp. 65–85.

Van der Vliet, N. (2010). Syntax-based discourse segmentation of Dutch text. In M. Slavkovik (Ed.), *Proceedings of the 15th Student Session*, ESSLLI, 2010, University of Copenhagen, pp. 203–210.

Van der Vliet, N., Berzlánovich, I., Bouma, G., Egg, G., Redeker, G.. (2011). Building a discourse-annotated Dutch text corpus. In S. Dipper & H. Zinsmeister (Eds.), *Beyond Semantics*, Bochumer Linguistische Arbeitsberichte 3, pp. 157–171.

Van der Vliet, N., Redeker, G. (2011). Explicit and implicit coherence relations in Dutch texts. *12th International Pragmatics Conference*. Manchester.

Voll, K., Taboada, M. (2007). Not all words are created equal. In *Proceedings of the 20th Australian Joint Conference on AI*, 2007.

Webber, B. (2009). Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of ACL-IJCNLP 2009*.

Wolf, F., Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31, pp. 249–287.