

Building a Discourse-Annotated Dutch Text Corpus

*Nynke van der Vliet**, *Ildikó Berzlánovich**, *Gosse Bouma**, *Markus Egg†*
and *Gisela Redeker**

* University of Groningen, †Humboldt University, Berlin.

Abstract

We are compiling a corpus of Dutch texts annotated with discourse structure and lexical cohesion, containing initially 80 texts from expository and persuasive genres. We are using this resource for corpus-based studies of discourse relations, discourse markers, cohesion, and genre differences. We are also exploring the possibilities of automatic text segmentation and semi-automatic discourse annotation. This paper discusses our design choices in text selection and segmentation and in the annotation of discourse structure and lexical cohesion.

1 Introduction

Discourse researchers from descriptive, cognitive, formal, and computational backgrounds unanimously subscribe to the view that texts are structured entities that exhibit coherence and cohesion (for a recent overview see Taboada and Mann (2006b)). Coherence refers to the way sentences or utterances combine to convey the informational and intentional (e.g., expressive or persuasive) meanings of the text. Cohesion refers to elements (conjunctions and other so-called “cue phrases”) that signal how utterances or larger text parts are related to each other, and to the way lexical elements like pronouns and definite noun phrases refer back to other items in the discourse (Halliday and Hasan, 1976). The main goal of our corpus-building effort is to provide the basis for investigating discourse structure, relational and lexical cohesion, and their interactions with genre, i.e., to support the modeling of textual organization.

Much of the theoretical and empirical research on relational coherence has focused on local coherence relations and their linguistic signaling (e.g., Sanders et al. (1992, 1993); Knott and Sanders (1998); Webber et al. (2003), Prasad et al. (2008)). Configurational issues concerning the hierarchical composition of larger stretches of text that arise from recursive application of coherence relations, have received some attention in computational linguistics, but lack a substantial empirical foundation. Various structures have been proposed, in particular, binary trees (e.g., Carlson et al. (2002); Stede (2004)), *n*-ary trees (e.g., Mann and Thompson (1988); Webber (2004), Polanyi et al. (2004); Thione et al. (2004)), and less constrained graph structures (Danlos (2004); Wolf and Gibson (2005)).

The interplay of relational discourse structure with referential and lexical cohesion has been investigated with a focus on the use and interpretation of anaphoric expressions (Fox (1987); Grosz et al. (1995); Kehler (2002); Poesio et al. (2004)); much less attention has been devoted to the role of lexical cohesion in co-determining the overall textual organization (but see Hasan (1984) and Hoey (1991)).

Textual organization cannot be studied without consideration of the variability between text genres (see, e.g., Eggins and Martin (1997), Webber (2009)). In particular, some texts are organized around a central purpose, e.g. a claim that is argued for or a request or proposal the text is intended to support, while descriptive or expository texts are usually organized around a central theme, moving through sub-themes or aspects. This difference is relevant for both, the relational structure and the role of lexical cohesion. The corpus therefore covers a range of genres.

By annotating relational and lexical organization in a variety of text types, this project will create a Dutch language resource for corpus-based discourse research, computational modeling, and applications like question answering and summarization.

2 Corpus design

Our aim is to provide a reliably annotated “gold standard” resource covering a range of genres. The emphasis on quality and richness of the manual annotation limits the size of the corpus, as careful annotation work is extremely time consuming.

2.1 Text selection

The corpus covers a range of text genres, including, in particular, expository texts, whose main purpose is to present information to the reader, and persuasive texts that aim to affect the readers intentions or actions. The texts vary in length between a minimum of approximately 190 words and a maximum of approximately 400 words. Longer texts become unwieldy for relational analysis, and top-level relations tend to be rather uninformative juxtapositions (Taboada and Mann, 2006b).

The corpus consists of 40 expository texts and 40 persuasive texts. For the expository subcorpus, 20 texts have been selected from online encyclopedias on astronomy¹ and 20 from a popular scientific news website². The persuasive texts are 20 fundraising letters from humanitarian organizations and 20 commercial advertisements from lifestyle and news magazines.

Encyclopedia entries as well as popular scientific news are learned exposition, i.e., texts that are strictly informational in purpose, but moderately technical in content and style, and that take the general public as their audience. In this way, we excluded scientific exposition, which is more abstract and technical in style and targets professional scientific audience (e.g., academic prose). Fundraising letters and advertisements are prototypical persuasive genres that have received much attention in the literature (e.g., Bhatia (1998), Kamalski (2007)). They have a clear and focused purpose and are directed at a general audience.

¹<http://www.astronomie.nl>; <http://www.sterrenwacht-mercurius.nl/encyclopedie.php5>

²<http://www.scientias.nl/category/astronomie>

2.2 Annotation

The starting point of our annotation work is a syntax-based segmentation of the texts into clausal atomic units, which has been developed in an extended training phase involving consistency checking aided by a collection of examples (see section 3 below). We then add annotations for discourse structure, relational cohesion, and lexical cohesion, which we are briefly introducing here (for details see sections 4 and 5).

For the analysis of relational discourse structures, we chose the widely used Rhetorical Structure Theory (RST) (Mann and Thompson, 1988; Taboada and Mann, 2006a) in its “extended classic” variant. The XML annotation is created using O’Donnell’s RSTTool³ (O’Donnell, 1997). The definitions of the RST relations are available from the RST website⁴.

Previous research has shown how combining genre analysis and RST analysis enriches our understanding of discourse structure (e.g., Taboada and Lavid (2003), Gruber and Muntigl (2005)). We are therefore overlaying the RST-trees with a segmentation of the global text units according to the genre-specific *moves* they realize (Upton and Cohen, 2009). The mapping of the sequence of moves onto the RST-trees adds relational and hierarchical information.

Three subsystems of cohesion contribute to the organization of a text: relational cohesion (lexical or phrasal elements that signal coherence relations), referential cohesion (anaphoric chains, spatial/temporal chaining and ellipsis), and lexical cohesion arising from the semantic network of the lexical items in the text (Halliday and Hasan (1976); Halliday and Matthiessen (2004)). In this project we focus on relational cohesion and lexical cohesion.

The analysis of relational cohesion will include all lexical or phrasal elements (discourse markers) in the text that signal coherence relations at local and global levels of discourse. We are currently developing our methodology for this analysis.

The analysis of lexical cohesion starts by identifying all content words (nouns, verbs, adjectives, adverbs) and then locating their neighboring lexical associates in other discourse units. The XML annotation is created with an MMAX-based tool (Müller and Strube, 2001).

All annotations are done separately by at least two annotators and then discussed. Inter-annotator agreement using Kappa shows a high level of agreement on the segmentation: .97 for the encyclopedia texts and .99 for the fundraising letters. We computed inter-annotator agreement for the RST analysis for two fundraising letters and two encyclopedia texts, using the methods proposed in Marcu et al. (1999). On average, the agreement was .88 on the spans and .82 on the nuclearity. The agreement on the RST relation labels was only .57. We suspect (and hope to confirm with the complete data set) that this is not a general deficiency of our annotation but a problem that can mainly

³available from <http://www.wagsoft.com/RSTTool/>.

⁴<http://www.sfu.ca/rst/>

be attributed to a few rather confusable relations such as *Joint* versus *Conjunction*. As Marcu et al. (1999) point out, these Kappa values are comparable with the agreement in other corpora.⁵

The annotation of all 80 texts in the core corpus will be complete by March 2011. Manuals detailing the segmentation and annotation rules will be made available along with the corpus.

3 Segmentation

An essential step in discourse analysis is the identification of suitable Elementary Discourse Units (EDUs). Various definitions of EDUs exist, ranging from a fine-grained segmentation to segmentation at sentence level. In classic Rhetorical Structure Theory (RST), clauses are considered to be EDUs, except for subject and object clauses, complement clauses, and restrictive relative clauses (Mann and Thompson, 1988).

For the annotation of the RST Discourse Tree Bank, Carlson and Marcu (2001) use a fine-grained segmentation in which they also treat complements of attribution verbs and phrases that begin with a strong discourse marker (e.g. *because of*, *in spite of*, *according to*) as separate EDUs. Relative clauses, nominal postmodifiers, or clauses that break up other legitimate EDUs are treated as embedded discourse units. Based on this, Lungen et al. (2006) developed segmentation guidelines for German text, but in contrast to Carlson and Marcu (2001) they exclude restrictive relative clauses, conditional clauses, and proportional clauses (clauses combined by comparative connectives). Grabski and Stede (2006) suggest to also include prepositional phrases as EDUs. Tofiloski et al. (2009) adhere more closely to the original RST proposals (Mann and Thompson, 1988) and segment coordinated clauses, adjunct clauses and non-restrictive relative clauses. To our mind, these differences follow from attempts to include semantic considerations in the definition of EDUs (i.e., including at least some proposition-denoting yet non-clausal segments among the EDUs).

For Dutch, as far as we know, such an elaborate investigation of what should count as an EDU has not yet been done. RST annotations of Dutch text have used the segmentation of the original RST proposals (Abelen et al., 1993) or taken clauses containing a finite verb (den Ouden et al., 1998) or whole sentences (Timmerman, 2007) as EDUs.

3.1 Segmentation principles

The segmentation we use for the Dutch corpus is fairly coarse. The EDUs are independent or subordinate clauses or other complete utterances (independent fragments). The definition of an elementary discourse unit is guided by the question of whether a discourse relation could hold between the unit and another segment. EDUs are typically

⁵Brown corpus (Francis and Kucera, 1979), MUC corpus (Chinchor, 2001), WSJ corpus (Carlson et al., 2002)

propositions or segments that constitute speech acts of their own. The segmentation principles are based on syntax and punctuation rather than semantic criteria.

Like Tofiloski et al. (2009), we treat simple sentences (1), coordinated clauses (2), subordinate clauses (3) and non-restrictive relative clauses (4) as EDUs.

- (1) [Elke donatie is waardevol!]
[Each donation is valuable!]
- (2) [Cavine kreeg aidsremmers][en dat maakte een levensgroot verschil.]
[Cavine got aids medication][and that made a huge difference.]
- (3) [Omdat de EU binnenkort beslist over nieuwe regels,][voeren we de druk op de politiek nu hoog op]
[Because the EU will decide on new regulations soon][we are now strongly increasing our pressure on politics.]
- (4) [Dit gat wordt veroorzaakt door een van de maantjes van Saturnus, Mimas,][die de ringen verstoort.]
[This gap is caused by one of the moons of Saturn, Mimas,][which disturbs the rings.]

In contrast to Tofiloski et al. (2009), we consider coordinated elliptical clauses (i.e. clauses that share a verb that is elided in one of the clauses, as in (5)) as separate EDUs, because the two clauses that share a verb can be seen as two separate predicates. This also applies to clauses that share a noun phrase as subject, as in (6). In Carlson and Marcu (2001), clauses with an ellipsed subject are segmented as EDUs as well, whereas clauses with an ellipsed verb are only treated as EDUs when there are strong rhetorical cues marking the discourse structure as in (7)⁶.

- (5) [De planeet draait in 58.6 dagen om haar as] [en in 88.0 dagen om de zon.]
[The planet turns around its axis in 58.6 days][and around the sun in 88.0 days.]
- (6) [De operatie duurde 15 minuten][en kostte 35 euro.]
[The surgery took 15 minutes][and cost 35 euros.]
- (7) [Back then, Mr. Pinter was *not only* the angry young playwright,] [*but also* the first] [to use silence and sentence fragments and menacing stares, almost to the exclusion] [of what we preciously understood to be theatrical dialog.]
(wsj_1936)

Non-restrictive relative clauses as in (8) and embedded clauses between parentheses as in (9) are considered to be embedded discourse units. Restrictive relative clauses, subject and object clauses, and complement clauses are not treated as separate EDUs (following classic RST). Contrary to Carlson and Marcu (2001), Lungen et al. (2006),

⁶Example from Carlson and Marcu (2001)

and Jasinskaja et al. (2007), we do not recognize non-clausal appositives as in (10) as separate EDUs.

- (8) [Echter gedurende de nacht, [die op Mercurius maanden lang kan duren,] daalt de temperatuur tot zo'n -185 graden Celsius.]
[However during the night, [which can last for months on Mercury,] the temperature decreases to about -185 degrees Celsius.]
- (9) [De binnenste maan [(van 2002 tot 2005 is dat Epithemeus)] beweegt iets sneller dan de buitenste] [en haalt die ander langzaam (met 450 meter per minuut) in.]
The innermost moon [(from 2002 to 2005 this is Epithemeus)] moves a bit faster than the outermost [and slowly (with 450 meters per minute) catches up with the other.]
- (10) [Het tweede type terrein, het laagland, telt relatief nog minder kraters dan het hoogland.]
[The second terrain type, the lowland, contains even fewer craters than the highland.]

Our segmentation uses punctuation in connection with syntax. Periods, exclamation marks and question marks are EDU boundaries, except for periods that are used in abbreviations, acronyms, dates and so forth. Independent fragments (subclausal expressions ending with a period) as in (11) are considered to be EDUs.

- (11) [Leuke hebbedingetjes.]
[Nice gadgets.]

Colon or semicolon are only treated as separation markers when the subsequent material is a clause as in (12). If it is a non-clausal expression, as in (13), it is not segmented. The same rule applies for text structures between hyphens or parentheses: clauses as in (9) or participle structures as in (14) are segmented as EDUs, but non-clausal material as in (15) is not segmented.

- (12) [Daar knapt ze zichtbaar van op;][ze begint ook weer te praten!]
[From that, she recuperates visibly;][she even starts to talk again!]
- (13) [In 2005 zijn nog twee maantjes van Pluto ontdekt: Nix en Hydra.]
[In 2005, two more small moons of Pluto were discovered: Nix and Hydra.]
- (14) [Wat er binnen deze bol [(horizon genoemd)] gebeurt weten we niet.]
[What happens inside this globe [(called horizon)] we don't know.]
- (15) [De krater Pan (inslagkrater), de grootste krater, is 100 kilometer in doorsnede] [en minstens 8 kilometer diep.]
[The crater Pan (impact crater), the biggest crater, is 100 kilometers in diameter][and at least 8 kilometers deep.]

4 Discourse structure

The annotation of discourse structure is intended to capture the hierarchical structures arising from coherence relations between discourse units, but also the genre-specific structures that can help in understanding genre differences in discourse structure.

4.1 Rhetorical Structure Theory

There is wide agreement that discourse is hierarchically structured, and many current theories assume that this structure arises from the recursive application of coherence relations. 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); see also the discussion in Taboada and Mann (2006a,b)) and we are researching such questions with our corpus. In addition, however, we are also interested in the configurational characteristics of discourse structure. We thus differ from annotation efforts like the Penn Discourse TreeBank (Prasad et al., 2008) that focus mainly on coherence relations and on implicit and explicit connectives. For us, it is essential to represent the full hierarchical structure of our texts.

Rhetorical Structure Theory (RST; Mann and Thompson (1988)) has proven successful for the analysis of whole texts and has been widely applied (for an overview see Taboada and Mann (2006a,b)) to texts of various languages and used for the annotation of large text corpora (Carlson et al. (2002), Stede (2004)).

We base our analyses on the set of 30 relations as defined in “extended classic” RST. We do not follow Carlson and Marcu (2001), who use a much larger set of relation labels (mostly necessitated by their more fine-grained segmentation) (for a critical discussion of both variants of RST, see Stede (2008)).

In particular, we do not use Carlson and Marcu’s (2001) *Attribution* and *Same* relations, which we consider problematic. *Attribution* is defined in Carlson and Marcu (2001) as the relation between a direct or indirect quotation and its attributing phrase or clause. This relation is arguably of a categorically different kind than coherence relations (Tofiloski et al. (2009), Skadhauge and Hardt (2005)). In classic RST, complement clauses and speech parentheticals are not considered as separate EDUs. This means that speech-reporting EDUs can enter coherence relations as speech events or by virtue of the speech that is reported (in particular when the quotation is continued in subsequent EDUs). This flexibility fits in well with the idea that semantic relations in discourse are often underspecified (Egg and Redeker, 2008).

The pseudo-relation *Same* is introduced by Carlson and Marcu (2001) to link two discontinuous parts of an EDU that is interrupted by another, parenthetically embedded, EDU. In classic RST, parenthetical EDUs are extracted and placed after their host EDU, thus obviating the need for a pseudo-relation (see, e.g., Redeker and Egg (2006)).⁷

⁷Borisova and Redeker (2010) point out problems involving the *Same* relation in the Discourse GraphBank (Wolf et al. (2003)).

4.1.1 Discourse trees or graphs?

Rhetorical Structure Theory assumes that the discourse structure of a text can be represented as an ordered tree. In this tree all text parts are in some way connected to the root of the tree, the most central text part. However, it has been claimed that tree structures are not sufficient to represent discourse structure (Asher (2008); Lee et al. (2008); Wolf and Gibson (2005)). Wolf and Gibson (2005) show that crossed dependencies (i.e. structures in which discourse units ABCD (not necessarily adjacent) have relations AC and BD) and multiple-parent structures (where a unit enters more than one coherence relation and is thus dominated by more than one node) occur abundantly in their Discourse GraphBank (Wolf et al. (2003)). They argue that these constellations, which violate the tree-structure constraints, are necessary to describe the text structures in their corpus, and that a more complex graph structure is thus required to represent the discourse structure of a text.

Webber (2006) and Egg and Redeker (2008, 2010), however, argue that the chain graphs in the Discourse GraphBank conflate discourse constituency and anaphoric dependency. Egg and Redeker (2008) point out that the analyses discussed in Wolf and Gibson (2005) have plausible tree-based alternatives and Egg and Redeker (2010) further support this argument with data from the Discourse GraphBank. While this question is not yet settled, we do find that trees are adequate data structures to represent the constituent structure of discourse for the texts in our corpus and thus use RST-trees to annotate discourse structures.

4.1.2 Non-binary trees

Given the assumption that discourse structure can be adequately represented by trees, it is tempting to consider the still stronger assumption that would only allow binary trees, which are much simpler and computationally more tractable. This restriction is indeed often implemented in discourse parsers (e.g. Marcu (2000); Soricut and Marcu (2003); Reitter (2003)). In our project, we choose plausibility and validity of our analyses over computational tractability and allow non-binary structures in our RST trees.

RST-trees do contain mostly binary relations (in particular the asymmetric *nucleus-satellite* relations),⁸ but they also admit non-binary structures with multiple nuclei or multiple satellites relating to one nucleus. In the first case, several nuclei are involved in one *multinuclear* relation, e.g., *List*, *Sequence* or *Joint*. Binary representations of such structures (proposed, e.g., by Egg and Redeker (2008)) involve a stacking of binary relations, implying a hierarchical ordering (left- or right-branching or pairwise clustering) among the list constituents. These binary representations do not reflect the

⁸RST distinguishes two kinds of relations: The asymmetric *mononuclear* relations like *Elaboration* or *Justify* relate a *nucleus* (centrally important) and a *satellite* (additional information, which could in many cases be left out without rendering the text incoherent). The symmetric *multinuclear* relations like *List* or *Joint* relate discourse entities of equal status.

equal importance of the items in the multinuclear relation.

In the second kind of non-binary structures, several *nucleus-satellite* relations share the same nucleus, e.g., when the central request of a fundraising letter is supported by various preceding or succeeding *Motivation* and/or *Justify* satellites, as described in Abelen et al. (1993), or when several separate *Elaborations* provide details about the contents of one nucleus. A binary representation of these structures requires that one of the satellites of the shared nucleus is included in the nucleus of another satellite, which is in many cases not plausible.⁹

We consider the regular occurrence of non-binary structures sufficient reason to assume that discourse structure representations require non-binary trees.

4.2 Moves

For comparisons of the global text structure across genres, we identify the genre-specific major building blocks of the texts using *move analysis* (Upton and Cohen, 2009). We identify the functional components, so-called moves, in the text. A move is realized by at least one EDU. Contrary to, e.g. Biber et al. (2007), we do not recognize moves below EDU level and do not allow embedding of moves. The moves in our analysis create a linear, non-hierarchical partition of the EDUs in the text. Each genre has a particular set of move types that occur regularly in texts of that genre. Some move types are obligatory. Any move type may be realized more than once in a particular text. In the encyclopedia entries, we identify the move types *name*, *define* and *describe*. For the fundraising letters, we follow Upton (2002), who identified seven move types labeled *get attention*, *introduce the cause and/or establish credentials of organization*, *solicit response*, *offer incentive*, *reference insert*, *express gratitude*, *conclude with pleasantries*. The move structure of advertisements is based on Bhatia (2005) and contains the following move types: *get attention*, *justify the product or service by establishing a niche*, *detail the product or service*, *establish credentials*, *endorsement/testimonial*, *offer incentive*, *use pressure tactics*, *solicit response*, and *reference to external material*. Finally, the starting point for determining the move structure of the popular scientific news will be van Dijk's superstructure of news (van Dijk, 1988), which is a hierarchical structure containing the main genre elements of news in general.

5 Cohesion

Parallel to the discourse structure annotation, we are annotating the corpus for relational cohesion and lexical cohesion.

⁹An alternative explanation that first collects all satellites in a *List* or *Joint* segment, which then as a whole functions as the sole satellite of the respective nucleus is only feasible in a subgroup of these cases, in which all satellites occur on the same side of the nucleus (before or after it) and are related to the nucleus in terms of the same relation.

5.1 Relational cohesion

Relational cohesion concerns the lexical or phrasal elements (*discourse markers*) in a text that signal coherence relations, both at the local and global levels of discourse. Some relations are often signaled by discourse markers, e.g. the conjunction relation (*and, also*), but others are implicit and do not contain clear cues (Taboada, 2006).

In a pilot study we have analyzed the distribution and explicit signaling of coherence relations in 20 encyclopedia entries and 20 fundraising letters. Intra-sentential relations are much more often signaled than inter-sentential relations (69% vs 16%), presumably reflecting the fact that intra-sentential clause combining usually involves an obligatory conjunction or adverb, while there is no such syntactic requirement for marking inter-sentential relations.

Future work will include the annotation of discourse markers (conjunctions and conjunctive adverbs) and their scopes, comparable to the annotations in the Penn Discourse Treebank (Prasad et al., 2008), with the dual aim of theoretical investigations and the development of a semi-automatic parsing tool for coherence relations.

5.2 Lexical cohesion

In our analysis of lexical cohesion, we aim to cover all types of semantic relations among lexical items in the text (see section 5.2.2 below; for recent work on an overview of approaches to lexical cohesion, see Tanskanen (2006)). We include only relations across elementary discourse units (EDUs), not within EDUs. This allows us to investigate the alignment between discourse structure and lexical cohesion, as both structures are based on the same units. At a finer level, we also study the co-occurrence of lexical cohesion types with coherence relations.

5.2.1 Selection of lexical items

As we are interested in the contribution of *lexical* cohesive relations, we exclude pronouns and do not follow referential chains through the text. The class of items for participating in lexical cohesion includes content words (nouns, verbs, adjectives, and adverbs of place, time, and frequency) and proper names. Proper names are treated as one unit. The elements of multi-word units (except for proper names) are treated as separate lexical items, while compounds are taken as indecomposable single units.

5.2.2 Categories of lexical cohesive relations

The categories we distinguish for lexical cohesive relations are listed in Table 1. By *repetition* we mean word repetition. The lexical items in full repetition have fully identical word form or they differ only in their inflectional suffix, whereas lexical items in partial repetition have different derivational suffixes in their word form. Under the heading

Category		Example
Repetition	Full repetition	<i>planet - planet</i>
	Partial repetition	<i>planet - planetary</i>
Systematic semantic relations	Hyponymy	<i>sun - star</i>
	Hyperonymy	<i>gas - hydrogen</i>
	Co-hyponymy	<i>Venus - Mercury</i>
	Meronymy	<i>planet - solar system</i>
	Holonymy	<i>solar system - sun</i>
	Co-meronymy	<i>Earth - sun</i>
	Synonymy	<i>life - existence</i>
Antonymy	<i>light - heavy</i>	
Collocation		<i>light - star</i>

Table 1: Categories of lexical cohesion

systematic semantic relations we include the traditional lexical semantic relations. The lexical cohesive relation *collocation* is formed between two lexical items which tend to occur in similar lexical environments because they describe things that tend to occur in similar situations or contexts in the world (Morris and Hirst, 1991). Note that this use of the term implies a meaning relation between the lexical items in contrast to its use in corpus linguistics, where collocation refers to the mere co-occurrence of words (Stubbs, 2001), which is not a sufficient criterion for lexical cohesion.

We identify relations arising from lexical meaning (e.g., *planet - Earth*) and ignore accidental meaning relations that arise from context. In addition, we identify relations that are easy for the reader to identify with general background knowledge and for which no further knowledge or textual context is necessary for their identification (e.g., we identify the relation of *astronomer* with *Kepler*, but not with *Richard Walker*, although the textual context helps us understand that Richard Walker is also an astronomer). This question is strongly related to the issue of register-sensitive and domain-sensitive relations. Although we aim to identify general relations, i.e., relations which are not specific of a certain register or domain, the annotators have to face the difficulties of drawing the line between general and context-dependent.

5.2.3 Lexical cohesion links as a graph structure

Lexical cohesive links build up graph structures in the text. In our analysis any candidate item can enter into a lexical cohesive relation with any other candidate items as long as there is a meaning relation between them. For each lexical item in a text, we identify its lexical links—if any—to preceding lexical items (lemmas), ignoring any links among the words inside an EDU. If a lexical item is linked to more than one preceding item, all of those relations are registered as cohesive links. Similarly, if a lexical item enters into cohesive relations with more than one item occurring in succeeding EDUs, all those links are counted.

In this way, we build up networks that represent the lexical cohesive structure of a text. By assigning graph structures to lexical cohesion, we differ from previous studies that identified lexical cohesive chains in text (e.g., Hasan (1984), Morris and Hirst (1991)) and follow those that identify networks (Hoey, 1991). Modeling lexical cohesion with graph structures provides a much richer representation than the lexical cohesive chains model. It also allows us to measure the centrality of a lexical item by its centrality in the network.

6 Conclusion

The resource we are building aims at a high standard of empirical validity (very careful annotation based on detailed, explicit rules) and coverage across a theoretically motivated selection of text genres. With a core of 80 texts, the corpus is rather small for computational applications, but still large enough for distributional analyses and structural comparisons.

We have been using the initially completed parts of this corpus to investigate genre differences in the use of discourse relations and in the occurrence of lexical cohesion relations and the interaction of these two aspects of textual organization (Berzlánovich and Redeker, 2011). As our discourse structure annotation follows the widely used “classic” RST, we expect our corpus to support cross-linguistic research through its comparability with RST-based corpora in other languages.

Our segmentation rules are surface oriented (based on syntax and punctuation) and have been implemented in an automatic segmenter (van der Vliet, 2010). Future work will include the annotation of discourse markers with the dual aim of theoretical investigations and the development of a semi-automatic parsing tool for coherence relations. With an eye on crosslinguistic research on discourse and discourse markers in the spirit of Knott and Sanders (1998), we will strive for compatibility with the annotation in the Penn Discourse TreeBank (Prasad et al., 2008), but will more freely allow markers to signal global coherence relations among larger text spans (which is discouraged by PDTB’s *Minimality Principle* (Prasad et al. (2007): 19), according to which annotators have to select the minimally necessary segments).

Finally, we also envisage combining our lexical cohesion analysis with computational coreference resolution (Hendrickx et al., 2008) and testing our network model of lexical cohesion against approaches based on lexical chaining (see, e.g., Barzilay and Elhadad (1997)).

Acknowledgments

The work reported here is supported by grant 360-70-280 of the Netherlands Organization for Scientific Research (NWO). For online documentation of the program *Modeling discourse organization* see www.let.rug.nl/mto. We are grateful to three anonymous reviewers for their valuable comments on an earlier version of this paper.

References

- Eric Abelen, Gisela Redeker, and Sandra A. Thompson. The rhetorical structure of US-American and Dutch fund-raising letters. *Text*, 13(3):323–350, 1993.
- Nicholas Asher. Troubles on the right frontier. In Peter Kühnlein and Anton Benz, editors, *Constraints in Discourse*. Benjamins, Amsterdam, 2008.
- Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, volume 17. Madrid, Spain, 1997.
- Ildikó Berzlánovich and Gisela Redeker. Genre-dependent interaction of coherence and lexical cohesion in written discourse. In *Corpus Linguistics and Linguistic Theory*, 2011. To appear.
- Vijay K. Bhatia. Generic patterns in fundraising discourse. *New directions for philanthropic fundraising*, (22):95–110, 1998.
- Vijay K. Bhatia. Generic patterns in promotional discourse. In Helana Halmari and Tuija Virtanen, editors, *Persuasion across genres: A linguistic approach*, pages 213–228. Benjamins, Amsterdam, 2005.
- Douglas Biber, Ulla Connor, and Thomas A. Upton. *Discourse on the move: Using corpus analysis to describe discourse structure*. Benjamins, Amsterdam, 2007.
- Irina Borisova and Gisela Redeker. Same and Elaboration relations in the Discourse Graphbank. In *Proceedings of the 11th annual SIGdial Meeting on Discourse and Dialogue, Tokyo*, 2010.
- Lynn Carlson and Daniel Marcu. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 2001.
- Lynn Carlson, Daniel Marcu, and Mary E. Okurowski. RST Discourse Treebank. *Linguistic Data Consortium*, 2002.
- Nancy Chinchor. *Message Understanding Conference (MUC) 7*. Linguistic Data Consortium, Philadelphia, 2001.
- Laurence Danlos. Discourse dependency structures as constrained DAGs. In M. Strube and C. Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Cambridge, Massachusetts, USA*, pages 127–135, 2004.
- Hanny J .N. den Ouden, Carel H. van Wijk, Jacques M.B. Terken, and Leo .G.M. Noordman. Reliability of discourse structure annotation. *IPO Annual Progress Report*, 33:129–138, 1998.
- Markus Egg and Gisela Redeker. Underspecified discourse representation. In P. Kühnlein and A. Benz, editors, *Constraints in Discourse (CID), Dortmund, June 3-5, 2005*, pages 117–138. Benjamins, Amsterdam, 2008.
- Markus Egg and Gisela Redeker. How complex is discourse structure? In *Proceedings of LREC'10, Malta, 17-23 May 2010*, pages 1619–1623, ELRA, 2010.
- Suzanne Eggins and Jim R. Martin. Genres and registers of discourse. In T.A. van Dijk, editor, *Discourse as Structure and Process*, volume 1, pages 230–257, 1997.
- Barbara A. Fox. *Discourse structure and anaphora: Written and conversational English*. Cambridge University Press, 1987.
- Nelson Francis and Henry Kucera. *Brown Corpus Manual*. Brown University, 1979.
- Michael Grabski and Manfred Stede. Bei: Intraclausal coherence relations illustrated with a German preposition. *Discourse Processes*, 41(2):195–219, 2006.
- Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.
- Helmut Gruber and Peter Muntigl. Generic and rhetorical structures of texts: Two sides of the same coin? *Folia Linguistica*, 39(1-2):75–113, 2005.
- Michael A.K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, 1976.
- Michael A.K. Halliday and Christian M.I.M. Matthiessen. *An introduction to functional grammar*. Arnold, London, 2004.
- Ruqaiya Hasan. Coherence and cohesive harmony. In J. Flood, editor, *Understanding reading comprehension: Cognition, language and the structure of prose*, pages 181–219. International Reading Association, Newark, 1984.

- Iris Hendrickx, Gosse Bouma, Frederik Coppens, Walter Daelemans, Veronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri van der Vloet, and Jean-Luc Verschelde. A coreference corpus and resolution system for Dutch. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, 28-30 May 2008*, 2008.
- Michael Hoey. *Patterns of lexis in text*. Oxford University Press, 1991.
- Katja Jasinskaja, Jörg Mayer, Jutta Boethke, Annika Neumann, Andreas Peldszus, and Kepa Rodríguez. Discourse tagging guidelines for German radio news and newspaper commentaries. Technical report, Universität Potsdam, 2007.
- Judith M.H. Kamalski. Coherence marking, comprehension and persuasion. On the processing and representation of discourse. *LOT Dissertation Series*, 158, 2007.
- Andrew Kehler. *Coherence, reference, and the theory of grammar*. Stanford, CA: CSLI, 2002.
- Alistair Knott and Ted Sanders. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30(2):135–175, 1998.
- Alan Lee, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Departures from tree structures in discourse: Shared arguments in the Penn Discourse Treebank. In *Proceedings of the Constraints in Discourse Workshop (CID08), Potsdam, Germany*, 2008.
- Harald Lungen, Csilla Puskàs, Maja Bärenfänger, Mirco Hilbert, and Henning Lobin. Discourse segmentation of German written text. In *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL 2006)*, 2006.
- William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.
- Daniel Marcu. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448, 2000.
- Daniel Marcu, Estibaliz Amorrortu, and Magdalena Romera. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL99 Workshop on Standards and Tools for Discourse Tagging*, pages 48–57, 1999.
- Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991.
- Christoph Müller and Michael Strube. 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*, pages 45–50, 2001.
- Michael O'Donnell. RST-Tool: An RST analysis tool. In *Proc. of the 6th European Workshop on Natural Language Generation, Duisburg*, 1997.
- Massimo Poesio, Rosemary Stevenson, Barbara D. Eugenio, and Janet Hitzeman. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363, 2004.
- Livia Polanyi, Martin van den Berg, Chris Culy, Gian L. Thione, and David Ahn. A rule based approach to discourse parsing. In *Proceedings of SIGDIAL '04. Boston, MA*, 2004.
- Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. The Penn Discourse TreeBank 2.0. Annotation manual. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania, 2007.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Language Resources and Evaluation*, 2008.
- Gisela Redeker and Markus Egg. Says who? On the treatment of speech attributions in discourse structure. In *Proceedings of Constraints in Discourse II*, pages 140–146, 2006.
- David Reitter. Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models. *LDV-Forum, GLDV Journal for Computational Linguistics and Language Technology*, 18:38–52, 2003.
- Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. Towards a taxonomy of coherence relations. *Cognitive Linguistics*, 15:1–35, 1992.
- Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. Coherence relations in a

- cognitive theory of discourse representation. *Journal of Pragmatics*, 4:93–133, 1993.
- Peter R. Skadhauge and Daniel Hardt. Syntactic identification of attribution in the RST treebank. In *Workshop On Linguistically Interpreted Corpora*, 2005.
- Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of HLT/NAACL 2003*, pages 228–235, 2003.
- Manfred Stede. The Potsdam Commentary Corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 96–102, 2004.
- Manfred Stede. Disambiguating rhetorical structure. *Research on Language & Computation*, 6(3): 311–332, 2008.
- Michael Stubbs. Computer-assisted text and corpus analysis: lexical cohesion and communicative competence. *The handbook of discourse analysis*, pages 54–75, 2001.
- Maite Taboada. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567–592, 2006.
- Maite Taboada and Julia Lavid. Rhetorical and thematic patterns in scheduling dialogues: A generic characterization. *Functions of Language*, 10(2):147–178, 2003.
- Maite Taboada and William C. Mann. Applications of rhetorical structure theory. *Discourse Studies*, 8(4):567–588, 2006a.
- Maite Taboada and William C. Mann. Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459, 2006b.
- Maite Taboada, Julian Brooke, and Manfred Stede. Genre-based paragraph classification for sentiment analysis. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 62–70, 2009.
- Sanna-Kaisa Tanskanen. *Collaborating towards coherence: Lexical cohesion in English discourse*. Benjamins, Amsterdam, 2006.
- Gian Lorenzo Thione, Martin van der Berg, Chris Culy, and Livia Polanyi. LiveTree: An integrated workbench for discourse processing. In B. Webber and D. Byron, editors, *ACL 2004 Workshop on Discourse Annotation, Barcelona, Spain*, pages 110–117, 2004.
- Sander E. J. Timmerman. Automatic recognition of structural relations in Dutch text. *MA thesis, University of Twente*, 2007.
- Milan Tofiloski, Julian Brooke, and Maite Taboada. A syntactic and lexical-based discourse segmenter. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 77–80, 2009.
- Thomas A. Upton. Understanding direct mail letters as a genre. *International Journal of Corpus Linguistics*, 7(1):65–85, 2002.
- Thomas A. Upton and Mary Ann Cohen. An approach to corpus-based discourse analysis: The move analysis as example. *Discourse Studies*, 11(5):585–605, 2009.
- Nynke van der Vliet. Syntax-based discourse segmentation of Dutch text. In Marija Slavkovic, editor, *Proceedings of the 15th Student Session, ESSLLI*, pages 203–210, 2010.
- Teun A. van Dijk. *News as discourse*. Erlbaum, Hillsdale, 1988.
- Bonnie Webber. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science*, 28(5):751–779, 2004.
- Bonnie Webber. Accounting for discourse relations: constituency and dependency. In M. Dalrymple, editor, *Intelligent linguistic architectures*, pages 339–360, 2006.
- Bonnie Webber. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 674–682, 2009.
- Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. Anaphora and discourse structure. *Computational Linguistics*, 29:545–587, 2003.
- Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287, 2005.
- Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. A procedure for collecting a database of texts annotated with coherence relations. Technical report, MIT, Cambridge, MA, 2003.