

# **Lexical cohesion and the organization of discourse**

First year report

PhD student: Ildikó Berzlánovich  
Supervisors: Prof. Dr. Gisela Redeker  
Dr. Markus Egg

Center for Language and Cognition Groningen  
University of Groningen

2008

## Table of contents

<b>1 Introduction.....</b>	<b>1</b>
<b>2 Lexical cohesion.....</b>	<b>2</b>
2.1 <i>Lexical cohesion and discourse organization.....</i>	2
2.1.1 Introduction.....	2
2.1.2 Lexical cohesion and genre.....	2
2.1.3 Lexical cohesion and coherence .....	3
2.2 <i>The role of lexical cohesion in the segmentation and centrality of discourse units.....</i>	5
2.2.1 Introduction.....	5
2.2.2 Discourse segmentation .....	6
2.2.3 Central discourse units.....	8
2.2.4 Conclusion .....	8
<b>3 Lexical cohesion analysis.....</b>	<b>9</b>
3.1 <i>The patterns of lexical cohesion: chains and networks .....</i>	9
3.2 <i>Classification of lexical cohesive relations .....</i>	12
3.3 <i>Elements forming lexical cohesive relations .....</i>	18
3.4 <i>The strength of lexical cohesive relations.....</i>	20
3.5 <i>Conclusion .....</i>	22
<b>4 Outline of the project.....</b>	<b>23</b>
4.1 <i>Research questions .....</i>	23
4.2 <i>Corpus.....</i>	24
4.3 <i>Methods.....</i>	24
<b>References.....</b>	<b>26</b>
<b>Appendix A: First year activities.....</b>	<b>30</b>
<b>Appendix B: Work plan .....</b>	<b>32</b>

## 1 Introduction

This report gives an outline of my PhD project *Lexical cohesion and the organization of discourse*. The PhD project is part of the NWO research program *Modelling Textual Organisation (MTO)*; <http://www.let.rug.nl/mto/>), which investigates the hierarchical organization of textual units into structured entities by coherence and lexical cohesion in different genres. The main goal of this PhD project is the multi-level analysis of discourse organization. More specifically, we study the interaction between coherence and lexical cohesion in persuasive and expository genres. The hierarchical structure of coherence has already been shown in previous studies (Mann & Thompson 1988). However, the hierarchical structuring of cohesion still needs to be developed. In order to explore the hierarchical structuring of discourse, two main organizing features are examined, viz., the *segmentation* and the *centrality* of discourse units.

This report focuses on one aspect of the PhD project, namely lexical cohesion (sections 2 and 3). First I briefly describe how lexical cohesion is connected to other discourse organizing properties (genre and coherence) (2.1). Then I discuss how lexical cohesion contributes to the organizing features *segmentation* and *centrality* of discourse units (2.2). I introduce the theoretical and methodological decisions for the lexical cohesion analysis in this project in section 3. Finally, section 4 provides an overview of the PhD project. Attached to this report are a list of my first year activities (Appendix A) and the work plan of my project for the next three years (Appendix B).

## 2 Lexical cohesion

### 2.1 Lexical cohesion and discourse organization

#### 2.1.1 Introduction

The *organization of discourse* is one of the central issues of discourse analysis. The term *organization* refers “to the sum of relations which hold between the units of text... and between each unit and the whole” (Goutsos 1997, p. 138). The term *discourse* refers to verbal communication in its situational and social context. When investigating the three levels of discourse organization (cohesion, coherence and genre), cohesion and coherence are analyzed in the individual texts. These texts belong to a certain genre, which places them into context.

Cohesion is thus one of the text properties that contributes to the organization of discourse. The term refers to the connectedness of the surface elements in the text. The three main categories of cohesion are referential cohesion (anaphoric chains), relational cohesion (connectives and ellipsis) and lexical cohesion. Lexical cohesion, which is the focus of this dissertation project, contributes to the ideational (semantic) structuring of discourse (Martin 1992). It refers to the semantic relations between the lexical items in the text; thus it provides information about the way lexemes are organized in the discourse (*lexical patterning*). See example (1) where lexical cohesive relations hold among the lexical items *sun*, *solar system*, *star*, *dwarf star* and *dwarf phase* in the text. (Note that our corpus contains Dutch texts, but all the examples have been translated into English in this report.)

- (1) After the forming of the **sun** and the **solar system**, our **star** began its long existence as a so-called **dwarf star**. 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.

The contribution of lexical cohesion to discourse organization can be captured better if we look at its interaction with other discourse-organizing properties with the aim to learn more about the systematic structuring of discourse (see the following sections 2.1.2 and 2.1.3).

#### 2.1.2 Lexical cohesion and genre

The concept of *genre* refers to the pragmatic knowledge shared by the members of a discourse community about a more or less conventionalized class of communicative events with common communicative purposes (cf. e.g., Swales 1990). This shared knowledge concerns standard default elements in texts of a particular genre, but also expectations about, e.g., subject matter and stylistic choices. With respect to discourse organization, we focus only on the former.

It has been emphasized since the early cohesion studies (e.g., in Halliday & Hasan 1976) that cohesion is sensitive to the varieties of discourse. Contrastive studies have shown that cohesion varies with the modality of discourse (i.e., spoken and written discourse) (Thompson 1994, Tanskanen 2006), with registers

(Louwerse, McCarthy, McNamara & Graesser 2004), and with spoken and written genres (Taboada 2004). Although lexical cohesion is present in the cohesion structure of all these forms of discourse, the distribution of the cohesive types strongly differs for genres. First of all, certain cohesive links occur more typically in certain varieties of discourse than others: referential cohesion is a characteristic type of narrative discourse when investigating participant chains (Fox 1987); ellipsis is typical of dialogical texts (Buitkiené 2005); conjunction is a favored cohesive link in the genres of academic discourse (Verikaitė 2005); finally, lexical cohesion is extremely dominant, for example, in the genres of legal discourse (Yankova 2006). The widely investigated genres in descriptive studies are narratives, which are a rich source for the analysis of participant chains, temporal and spatial progression. Most computational linguistic studies analyze news documents – for its free accessibility and for users' demand for tools to manage the constantly growing data of news (Stokes 2004).

In our research we focus on expository and persuasive genres. Information-oriented expository texts present facts. The units of discourse are formed around related concepts, topics and their subtopics. Each discourse unit of the expository texts aims to elaborate on the introduced topic or to move on to a subtopic or a new topic in order to provide new information about the main topic for the reader (Britton 1994). The ideational structure thus seems to be dominant in expository texts. As the linear organization of text follows the clustering of information, we assume that lexical cohesion built upon semantic relations reflects the clusters. While the emphasis is on the content in expository texts, persuasive texts are built around a central illocutionary force (i.e., 'persuade') to have an effect on the reader. Persuasive texts present material that buttresses this strong illocutionary force. The discourse units support the central illocution, they are expected to rely less on lexical resources. Hence, the rhetorical structure of discourse dominates over the ideational structure in persuasive texts. We hypothesize that lexical cohesion (contributing to the ideational structure of discourse) plays a pivotal role in the structuring of expository texts, whereas it is less prominent in the case of persuasive texts.

### ***2.1.3 Lexical cohesion and coherence***

There are two main views on coherence. One regards it as a property of text, focusing on the formal criteria that distinguish texts from non-texts. Another approach views coherence as a discourse processing concept (Hellman 1995). The coherence of a text arises from the processes of text production and comprehension (e.g., Sanders & Noordman 2000). Our approach to coherence analysis follows Rhetorical Structure Theory (RST) (Mann & Thompson 1988) which has empirically proved to be useful for the study of coherence in different languages and different genres. It is process-oriented in the sense that the relations between the discourse units are defined in terms of the writer's purposes (based on the analysts' plausibility judgements and semantic criteria in the relation definitions).

Coherence has to be clearly distinguished from cohesion. Cohesion refers to the overt semantic relations in the text, whereas coherence refers to semantic and pragmatic relations between text parts which are interpretable against the background of specific world knowledge (de Beaugrande & Dressler 1981, Enkvist 1990). It has been widely discussed whether both cohesion and coherence are necessary for the organization of discourse. It has been argued that cohesion is a necessary, but not sufficient criterion of coherence (e.g., Halliday & Hasan 1976, Halliday 1985). It has also been claimed that cohesion is neither necessary, nor sufficient for the coherence of a text, and a text can be coherent without formal cohesive devices (Hoey 1991, Hellman 1995). As there is evidence for the relevance of both cohesion and coherence in text, we take both cohesion and coherence as contributing to discourse organization: cohesion being at the surface level of the text, whereas coherence being an underlying phenomenon in the text. Since their role in discourse organization is genre-dependent, in certain genres cohesion, in other genres coherence might be more dominant in the organization of discourse.

The coherence structure of a text can be captured by looking at the relations between the text parts. Coherence relations are classified into subject-matter (semantic or ideational) and presentational (pragmatic or interpersonal) relations depending on the source of the relation (Taboada & Mann 2006). Semantic relations result from the locutionary meanings of the text parts. See example (2) where the second text part provides more detailed information about the content of the first text part.

- (2) [After the forming of the sun and the solar system, our star began its long existence as a so-called dwarf star.] [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.]

The first text part introduces the topic ‘sun as a dwarf star’, and the second text part presents details of what it means for the sun to be a dwarf star. The coherence relation between the text parts is called *Elaboration*. Other examples of semantic relations are *Cause*, *Circumstance* and *Interpretation*.

In contrast, pragmatic relations arise from the illocutionary meanings of the text parts. This is illustrated in example (3) where the first text part justifies the writer’s right to present the request in the second text part.

- (3) [In this brochure you can read how much the sick children enjoy the visit of the CliniClowns.] [Fill in the giro form, and support the work of the CliniClowns.]

Drawing attention to the attached brochure (first text part) makes the reader more ready to accept the request for financial support (second text part). This relation is called *Justify* in the RST tradition. The set of pragmatic relations includes, for instance, *Motivation*, *Evidence* and *Concession*.

However, it has been pointed out that the clear distinction between semantic and pragmatic relations is problematic. Just as there is often an ideational coherence relation underlying and enabling a more salient pragmatic one (Redeker 2000), lexical cohesion primarily contributing to the ideational

organization of discourse may also indirectly contribute to the interpretation of pragmatic relations. When looking at the interaction between coherence and lexical cohesion, we gain insight into the role of lexical cohesion in contributing to the locutionary and illocutionary meaning between text parts (i.e., subject matter and presentational coherence).

In our project we study the interaction between coherence and lexical cohesion in detail. Investigations of the interrelation between coherence relations and lexical cohesion have found that, for instance, the *Elaboration* relation is frequently established by relations like hyperonymy/hyponymy and holonymy/meronymy (Bärenfänger, Lobin, Lungen & Hilbert 2006).

## **2.2 The role of lexical cohesion in the segmentation and centrality of discourse units**

### **2.2.1 Introduction**

Cohesion analysis has gained much attention in several branches of linguistics. Most descriptive studies (Halliday & Hasan 1976, Hasan 1984, Halliday 1985, Hoey 1991, Martin 1992, Halliday & Matthiessen 2004, Tanskanen 2006) aim to develop an appropriate taxonomy for the analysis of all kinds of texts. In order to find a suitable categorization and to generalize the results, a large amount of data is necessary. This has led to the increased use of computerized text corpora in linguistic research since the late 1980s (Conrad 2002).

The analysis of large corpora demanded the development of computational tools. Online lexical databases (e.g., WordNet – Fellbaum 1998) and discourse parsers (e.g., Polanyi, Culy, van den Berg, Thione & Ahn 2004) have been developed. These tools have been used for research on natural language processing focusing on different textual phenomena. As lexical cohesion is relatively easy to compute compared to other text properties (for example, coherence), it has been widely investigated in computational linguistics. Research on lexical chaining (Morris & Hirst 1991), coreference resolution (the COREA project, <http://www.cnts.ua.ac.be/~hoste/corea.html>) and topic detection and tracking (Stokes 2004) are all related to the study of lexical cohesion. Two common computational applications of these analyses are automatic text summarization (e.g., Silber & McCoy 2002) and thematic segmentation (e.g., Ferret 2007). These applications make use of the information on two crucial phenomena in the hierarchical organization of discourse: discourse segmentation and the centrality of certain discourse units compared to less central text parts. The main questions of the following two sections are:

- How does lexical cohesion contribute to the segmentation of discourse?
- How does lexical cohesion contribute to the centrality of certain discourse units?

### 2.2.2 Discourse segmentation

Discourse segmentation means the identification of the boundaries between shorter and longer stretches of discourse. It concerns the identification of the so-called elementary discourse units (EDU), which provides the base for the analysis of both cohesion and coherence. The EDU is defined in different ways in different studies. Looking at intersentential cohesive links, Halliday and Hasan (1976) take sentence as the unit of analysis. This has been adopted in many following studies (e.g., Hoey 1991, Tanskanen 2006). Halliday (1985) suggests the ‘clause complex’, Halliday and Matthiessen (2004) the clause as the unit of analysis instead of the sentence. Tanskanen (2006, p. 84) argues for her choice of the *sentence* as the analytical unit for written discourse and prepared spoken discourse (i.e., speeches) and the *turn* for conversations in order to “overcome a potential written-language bias in studies of spoken language.” A strong practical argument from a computational linguistic viewpoint is that the use of discourse parsers for the segmentation of sentences into smaller units of analysis is costly (Barzilay & Elhadad 1999).

Many computational studies (e.g., Hirst & St-Onge 1998) include both intrasentential and intraclausal relations when analyzing cohesion. Intrasentential relations are regarded as cohesive links even in the descriptive approach of Tanskanen (2006, p. 85), because “they may help to make the unity of a sentence clearer”. They ignore the fact that cohesion as a property of texts should be investigated across EDUs. As cohesion is a discourse phenomenon, the cohesive relations link two EDUs together (e.g., Hoey 1991).

Following the RST tradition (Taboada & Mann 2006) we take the *clause* as the smallest unit of discourse (EDU). The boundaries between the EDUs are thus clause boundaries. This view is followed in our analysis of coherence and cohesion. Furthermore, we consider cohesion as a text property. The semantic relations within the EDU, hence, fall outside the scope of the lexical cohesion analysis in this paper. (For problematic cases and the complexity of discourse segmentation see Korfiatis 2007.)

Besides EDUs, discourse segmentation deals with longer stretches of discourse as well. It has been shown that lexical cohesion might be an indicator of discourse segmentation. Morris and Hirst (1991) found a close correspondence between lexical chains and structural unit boundaries. Lexical chains in Morris and Hirst (1991) are built up of lexical items linked on the basis of the relations in a thesaurus (i.e., prominently with traditional semantic relations). Ferret (2002) investigates the role of collocations when thematically segmenting texts. They both point out that topic boundaries can be indicated when analyzing lexical cohesion. We assume that topic shifts may be detected in the lexical cohesive structure of information-oriented expository texts, but less so in strongly intentional persuasive texts.

Paragraph is a discourse unit usually built up of more than one EDU. Paragraph segmentation may be strongly related to topic segmentation in certain texts. Filippova and Strube (2006, p. 268) point out that “if there is a topic boundary, it is very likely that it coincides with a paragraph boundary. However, the reverse is not true and one topic can extend over several paragraphs.” The

identification of paragraphs thus seems to be challenging for computational applications. Filippova and Strube (2006) argue that paragraphing as a stylistic phenomenon needs more criteria (e.g., discourse markers, pronouns, information structure) for its identification, and besides lexical elements function words should be investigated as well. Hence, low cohesion as a signal of paragraph boundaries (Bolshakov & Gelbukh 2001) might be questioned.

It has been shown in psycholinguistic studies that readers can intuitively identify paragraph boundaries (Hoey 2005). The investigation of paragraphing thus remains an interesting issue. The question to what extent lexical cohesion reflects paragraphing needs more investigation and is a main concern of the present project. It is assumed at this point that it might be a reliable indicator of paragraph boundaries in genres where lexical cohesion is a dominant organizing feature (e.g., expository texts) in contrast to genres where either other types of cohesion occur in greater proportion, or other text properties play an important role. For genres where lexical cohesion is dominant, it is assumed that lexical cohesion will reflect topic shifts in the discourse as well. If lexical cohesion has a less central role in discourse organization in certain genres, it cannot be regarded as a strong predictor of discourse segmentation. (For the differences between the structuring of expository and persuasive texts see section 2.1.2.)

For higher levels of discourse (i.e., for larger discourse units) it has also been investigated how various surface markers (e.g., adverbials of time and place, connectives, punctuation) in the text signal discourse segmentation. The use of such surface markers is especially related to topic shifts. Experimental studies have pointed out that these segmentation markers strongly influence discourse comprehension and production (Bestgen 1998, Bestgen & Vonk 2000). It has to be emphasized again that this project focuses on the semantic relations of content words. Segmentation markers fall outside the scope of this research, as they are more related to relational and referential cohesion, but not to lexical cohesion.

So far we have looked at the segmentation of particular texts at lower levels (EDUs) and at higher levels (larger discourse units) of discourse organization. At a more global level of discourse organization we examine the genre-specific structure of the texts. One approach towards genres investigates the global structure of different genres, the so-called *move structure* (Swales 1990). *Moves* are the functional components of the genre structure, each of them contributing to the main communicative purpose of the given genre. It has to be noted that not all the moves are realized in the texts, or certain moves might be realized more than once in a text. Thus, their order varies in the particular texts. As the length of the discourse units realizing the moves differs as well (from clauses to longer stretches of discourse), the automatic segmentation of the moves seems problematic. In our project the move structure is mapped onto the top level of the coherence structure of the particular texts, which is then compared to the structuring of lexical cohesion (for details see the pilot study in Berzlánovich, Egg & Redeker 2008).

### 2.2.3 Central discourse units

When looking at the hierarchy in discourse organization, the discourse units defined by the segmentation are compared in terms of their centrality in discourse. In the coherence structure more and less central discourse units can be distinguished. In the RST tradition, the units that are most central to the writer's purposes are called the *nuclei*; less central supporting or expanding units are called *satellites* (Mann & Thompson 1988). We assume that lexical cohesion not simply reflects the boundaries between discourse units, but these discourse units can be ranked according to their centrality in the discourse organization. Similarly to the coherence structure, we can find more and less central discourse units in the structure of lexical cohesion, depending on the extent to which they contribute to building the lexical cohesive structure in a text.

Centrality with regard to cohesion is addressed in Hasan (1984). When looking at the interaction of cohesive chains in the texts, she distinguishes central tokens (the shared items of chains participating in the interaction) and peripheral tokens (items not participating in chain forming). While Hasan (1984) investigates chain interaction and central chain members, Hoey (1991) is concerned with the centrality of sentences within the text. In his analysis of lexical cohesive networks, sentences with the highest level of bonding (i.e. with the highest number of cohesive links) form the most central parts of the text compared to the marginal sentences that are lexically not bonded with other sentences. With this division he shows that either the elimination of marginal sentences or the selection of the central ones provides a summary of the analyzed text. (More about lexical chains and networks in section 3.1.)

The centrality of lexical items and discourse units provides the base for text summarization, a major application of lexical cohesion in computational linguistics. Barzilay and Elhadad (1999) argue that the identification and extraction of the strong lexical chains representing lexical cohesion gives the summary of the analyzed text. The general idea in Silber and McCoy (2002) for text summarization is similar. Measuring the centrality of discourse units in lexical cohesion is still an unexplored area. It is thus one of the main challenges of this project.

### 2.2.4 Conclusion

- Both for coherence and lexical cohesion analysis the elementary discourse units are clauses.
- Semantic relations within an EDU are ignored in lexical cohesion analysis.
- Both coherence and lexical cohesion are investigated how they contribute to discourse segmentation.
- The realization of the functional components (moves) of the genre structure differs in the texts.
- Both coherence and lexical cohesion are investigated how they contribute to the centrality of discourse units.
- The measurement of centrality in lexical cohesion has to be developed.

### 3 Lexical cohesion analysis

The ways to investigate the boundaries between discourse units and to identify the central text parts vary in the studies discussed above. The main differences center around the following problems. Cohesion can be viewed as forming a pattern of chains or a network. We need to consider which patterning is suitable in the case of lexical cohesion. The criteria for the identification of lexical cohesive relations have to be defined as well. Then the elements forming lexical cohesive relations have to be described. Finally, stronger and weaker lexical cohesive relations need to be distinguished in order to identify central discourse units and to see how the hierarchy of discourse organization in the case of lexical cohesion is built up.

#### 3.1 The patterns of lexical cohesion: chains and networks

There is a strong division in the literature when looking at the patterning of lexical cohesion. It might be interpreted as *chains* or as a *network* organizing the related items into a sequence (chain) or a net (network) in the text. Both views are sensible in the interpretation of the patterning of lexical cohesion. The lexical items are semantically related to other lexical items in the text; hence, they form a network. At the same time, certain nominal items, for instance, may have identity of reference with other items in the text (hence, referential chains).

The genre of the analyzed text also determines if the patterning of cohesion is dominantly chaining or a network. Texts of certain genres (e.g., non-narrative or legal texts) where lexical cohesion is overwhelming compared to other cohesive relations, it is reasonable to analyze nets of cohesive links. However, in texts containing many anaphoric links, temporal and spatial progression (e.g., narratives) or genres with many relational cohesive links (e.g., written genres in academic discourse) the chaining aspect of cohesion gains more emphasis.

Hasan (1984) argues that the overall cohesive structure of a text is captured best when looking at the so-called *cohesive chains* and their interaction through the text. With introducing the idea of the *cohesive chain* she proposes a distinction between two types of chains: identity chain based on co-referentiality and similarity chain built upon non-textbound semantic relations.

The chaining approach always takes the last preceding item of the chain as the antecedent of the following element to make a cohesive relation. Example (4) illustrates this method. (Lexical chain candidates are italicized; the items entering a chain are bolded; the index numbers stand for the identification of the chains; *EDU* is for *elementary discourse unit*.)

- (4) EDU<sub>5</sub>[After the ***forming***<sub>6</sub> of the ***sun***<sub>1,3,4</sub> and the ***solar system***<sub>4</sub>, our ***star***<sub>3,4</sub> began ***its***<sub>3</sub> long ***existence***<sub>5</sub> as a so-called ***dwarf star***<sub>4</sub>.] EDU<sub>6</sub>[In the ***dwarf phase***<sub>4,6</sub> of ***its***<sub>3</sub> ***life***<sub>5,6</sub>, the ***energy*** that the ***sun***<sub>1,3,4</sub> ***gives off*** is generated in ***its***<sub>3</sub> ***core*** through the ***fusion*** of ***hydrogen*** into ***helium***.] EDU<sub>7</sub>[The ***sun***<sub>1,3,4</sub> is about five billion ***years***<sub>2</sub> ***old now***] EDU<sub>8</sub>[and ***it***<sub>3</sub> still has enough ***fuel*** for another five billion ***years***<sub>2</sub>.]

Here we find repetition chains (CHAIN 1:  $sun_{EDU5} - sun_{EDU6} - sun_{EDU7}$ ; CHAIN 2:  $years_{EDU7} - years_{EDU8}$ ), referential chains (CHAIN 3:  $sun_{EDU5} - star_{EDU5} - its_{EDU5} - its_{EDU6} - sun_{EDU6} - its_{EDU6} - sun_{EDU7} - it_{EDU8}$ ), chains arising from traditional semantic relations (e.g., hyponymy, meronymy, synonymy; CHAIN 4:  $sun_{EDU5} - solar\ system_{EDU5} - star_{EDU5} - dwarf\ star_{EDU5} - dwarf\ phase_{EDU6} - sun_{EDU6} - sun_{EDU7}$ ; CHAIN 5:  $existence_{EDU5} - life_{EDU6}$ ; CHAIN 6:  $forming_{EDU5} - dwarf\ phase_{EDU6} - life_{EDU6}$ ) (note that the boundaries between EDUs are ignored in this example).

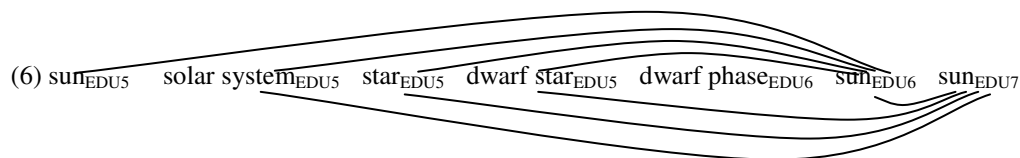
Many computational analyses (e.g., Morris & Hirst 1991, Silber & McCoy 2002) work with lexical chains to compute lexical cohesion. The relations within these lexical chains are defined with semantic networks, mostly machine-readable dictionaries, thesauri, or the WordNet database. The studies on lexical chains face the problem that chains may share lexical items. See example (4) where the items  $sun_{EDU5}$ ,  $star_{EDU5}$ ,  $sun_{EDU6}$ ,  $dwarf\ phase_{EDU6}$ ,  $life_{EDU6}$  and  $sun_{EDU7}$  occur in more than one chain. The phenomenon where a lexical item of one chain reoccurs in another chain is called *chain return* in Morris and Hirst (1991). They realize that merging lexical chains is unavoidable when analyzing lexical cohesion. Morris and Hirst's (1991) *chain return* is different from what Hasan (1984) calls the interaction of cohesive chains. Hasan (1984) links chains together when the items of different chains have an identical function relation, whereas Morris and Hirst (1991, p. 35) merge chains when the lexical items are shared "as a result of the intentional merging of ideas or concepts in the text". Still, the computational analyses prefer keeping lexical chains separate for the sake of easier chain computation. To solve the problem of merging chains and to avoid chain returns (shared lexical items), word distance constraints are defined for the relations building up the chains (e.g., Barzilay & Elhadad 1999, Stokes 2004).

We consider lexical cohesion as a network of relations. This is in accordance with our definition of lexical cohesion, i.e. looking at the semantic relations of lexical items in the text. If one item forms semantic relations with more than one lexical element in the text, all the relations are regarded as lexically cohesive. Compare example (4) with example (5). While example (4) shows the lexical chains, example (5) highlights all the lexical cohesive relations of each lexical item. (Candidate lexemes are italicized; the items semantically related to (an) earlier item(s) are bolded; the relation types and the related items are in index; *EDU* is for *elementary discourse unit*.)

- (5)  $EDU_5$ [After the *forming* of the ***sun***<sub>repetition(sunEDU4),hyponymy(starEDU4),co-hyponymy(Proxima Centauri EDU4),co-meronymy(EarthEDU3)</sub> and the ***solar system***<sub>holonymy(sunEDU4),holonymy(starEDU4), holonymy (ProximaCentauriEDU4), holonymy(EarthEDU3)</sub>, our ***star***<sub>repetition(starEDU4),co-meronymy(EarthEDU3), hyperonymy (sunEDU4),hyperonymy(Proxima CentauriEDU4)</sub> began its long *existence* as a so-called ***dwarf star***<sub>hyperonymy(sunEDU5),hyponymy(starEDU4),co-meronymy(ProximaCentauriEDU4),co-meronymy(EarthEDU3).</sub>]  
 $EDU_6$ [In the ***dwarf phase***<sub>co-meronymy(formingEDU5), collocation(dwarf starEDU5)</sub> of its ***life***<sub>meronymy(formingEDU5),synonymy(existenceEDU5)</sub>, the *energy* that the ***sun***<sub>repetition(sunEDU5),hyponymy(starEDU5), hyponymy (dwarf starEDU5),co-meronymy(EarthEDU3),meronymy(solar systemEDU5),co-meronymy(Proxima CentauriEDU4)</sub> gives off is *generated* in its core through the *fusion* of *hydrogen* into *helium*.]  
 $EDU_7$ [The ***sun***<sub>repetition(sunEDU6),hyponymy(starEDU5),co-meronymy(EarthEDU3),co-hyponymy(ProximaCentauriEDU4),meronymy(solar system EDU5),hypo-nymy(dwarf starEDU5)</sub> is about five billion ***years***<sub>repetition(yearEDU4)co-</sub>

meronymy(minutes<sub>EDU3</sub>),co-meronymy(months<sub>EDU4</sub>) old now] <sub>EDU8</sub>[and it still has enough fuel for another five billion years repetition(year<sub>EDU7</sub>),co-meronymy(minutes<sub>EDU3</sub>),co-meronymy(months<sub>EDU4</sub>)]

The comparison of example (4) and example (5) shows that when looking for lexical chains in the text, many semantic relations may remain undetected. For instance, CHAIN 6 reflects the lexical cohesive force between the lexical items *forming*<sub>EDU5</sub>, *dwarf phase*<sub>EDU6</sub> and *life*<sub>EDU6</sub>. However, it does not clearly show the relations between *forming*<sub>EDU5</sub> – *dwarf phase*<sub>EDU6</sub> and *forming*<sub>EDU5</sub> – *life*<sub>EDU6</sub>. With this, the more central role of the item *forming*<sub>EDU5</sub> in the chain also remains hidden. With assigning the pattern of a network to lexical cohesion, all the lexical cohesive relations per lexical item can easily be traced. Take the longer example of CHAIN 4 in example (4), containing seven lexical items. The chain only shows which lexical items are involved in traditional semantic relations. Moreover, it might give the misleading impression that the chain is built by linking one item to the succeeding one, as in the case of the repetition chain (CHAIN 1 and CHAIN 2) and the referential chain (CHAIN 3). The difference between the nature of CHAINS 1 through 3 and CHAIN 4 results from the fact that the first three chains are transitive, and CHAIN 4 is intransitive. CHAIN 4 does not reflect which item is related to which item. With this, the particular relations are not presented, and it also remains unclear which items play a major role in building up the chain. Example (6) shows the lexical items in CHAIN 4, now highlighting all the lexical cohesive links built up of traditional semantic relations.



Example (6) shows the central role of the lexical items *sun*<sub>EDU6</sub> and *sun*<sub>EDU7</sub> participating in both shorter and longer distance lexical cohesive relations.

Example (5) presents all this information. It clearly illustrates that the lexical items tend to form more than one semantic link with previous items. This is what we mean by the network of lexical cohesion. Viewing lexical cohesion as a network also enables us to identify more and less central lexical items, which is relevant information to detect the cohesion structure.

We claim that the analysis of chains is not sufficient for investigating lexical cohesion. As Hoey (1991, p. 10) has already pointed out, “lexical cohesion is the only type of cohesion that regularly forms multiple relationships.” These “multiple relationships” form networks spreading across the text. In other words, referential and certain elements (e.g., signals of temporal progression) of relational cohesion are linear, whereas lexical cohesion is non-linear in the discourse structure. Experiments with readers (Beigman Klebanov & Shamir 2006) have led to the same conclusion. Readers intuitively tend to create multiple relations between lexical items in the text; the overall lexical cohesion is thus not a binary phenomenon. These relations can be represented by graphs. In short: “Models based on mutually exclusive lexical chains will not suffice” (Beigman

Klebanov & Shamir 2006, p. 109). We strongly agree with this statement, and take it as a powerful motivation to regard lexical cohesion as a network phenomenon in the text.

However, example (5) shows that even within the network approach we imposed certain constraints. In the semantic relations we decided always to look at the last occurrence of the preceding item that forms a relation with the item under consideration. Note, for example, that *sun*<sub>EDU6</sub> is linked with a hyponymic relation to *star*<sub>EDU5</sub>, but not to *star*<sub>EDU4</sub>. This restriction is necessary in order to avoid chaotic, overloaded networks. With this procedure, in the case of repetition relations cohesion is created in the pattern of transitive repetition chains. See, for instance, the repetition chain *sun*<sub>EDU7</sub> – *sun*<sub>EDU6</sub> – *sun*<sub>EDU5</sub> – *sun*<sub>EDU4</sub>: *sun*<sub>EDU7</sub> forms a repetition relation with *sun*<sub>EDU6</sub>, but not with *sun*<sub>EDU5</sub> or *sun*<sub>EDU4</sub>; *sun*<sub>EDU6</sub> forms a repetition relation with *sun*<sub>EDU5</sub>, but not with *sun*<sub>EDU4</sub>, etc. In contrast, the transitivity of traditional semantic relations (for instance, hyponymy, meronymy and synonymy) is hard to follow. While these systematic semantic relations build up a semantic network, we talk about chains for word repetition. However, it does not mean that repetition does not enter into the cohesion network. Note the example of the item *sun*<sub>EDU7</sub> that forms a repetition link with *sun*<sub>EDU6</sub>, and at the same time it is also linked to other lexical items (*star*<sub>EDU5</sub>, *Earth*<sub>EDU3</sub>, *Proxima Centauri*<sub>EDU4</sub>, *solar system*<sub>EDU5</sub> and *dwarf star*<sub>EDU6</sub>). (Further differences between lexical cohesive relations are discussed in more details later in section 3.2.)

In sum, we do not deny that cohesive chains can be detected in the text. They clearly play an important role in cohesion. Good examples are anaphoric chains or the temporal and spatial progressions in narratives. What we suggest is that certain cohesive types (referential cohesion and elements of relational cohesion such as coordinatives or markers of temporal and spatial progression) can indeed be usefully investigated as chains in the cohesion structure. A cohesive item of the referential cohesion always points backward or forward to another specific item with identity of reference. Coordinatives in relational cohesion create a cohesive link between two successive elementary discourse units (with grammatical terms, clauses). Markers of time and place signal the linear progression in narratives. Lexical cohesion, by contrast, should be investigated as a network of relations.

### 3.2 Classification of lexical cohesive relations

There have been many attempts to find the appropriate classes to describe cohesion. In this section the categorization is introduced which we follow in our analysis of lexical cohesion. Although it has been influenced by previous taxonomies, we do not discuss them here one by one. Instead, we provide information about other categorizations while discussing the categories used in the project (for a summary of the previous approaches see Tanskanen 2006).

Previous studies show how difficult it is to identify the lexical cohesive relations and to provide an appropriate categorization for the analysis of lexical cohesion. A recent approach to solve these problems is the reader-oriented, experimental approach. These experiments aim to explore lexical cohesion as it is

perceived by the readers. In Morris and Hirst (2004) the readers were asked to find words in the text that were related to each other in meaning. There was 63% agreement on the identification of the related word groups and less for word pairs when readers were asked to give labels to the relations. The importance of these studies is that they show that meaning relations between the words in the text are recognized by the readers, and they can classify these relations by their meanings. Most of the identified word pairs were collocations in the experiments. (Note that these are the results of the analysis of a text with a certain genre, namely a general-interest article. The findings can thus not be generalized.) These studies emphasize the relevance of the category of *collocation* (non-systematic semantic relations) in the lexical cohesion taxonomy. Furthermore, the labeling of the relations might contribute to the refinement of the further categorization of collocation. Finally, psychological lexical-priming experiments have investigated the relation between lexical elements. It has been found that not only traditional semantic relations, but also collocation relations are perceived by the readers (Hodgson 1991).

Another remark on the identification of lexical cohesive relations concerns genre. Similar to the general observation that cohesion varies with genre, lexical cohesion tends to vary with genre as well. Conclusions of previous studies claiming that non-systematic relations are overwhelming compared to systematic relations (Morris & Hirst 2004) are precipitous. The proportion of the lexical relations in the cohesive pattern in a text strongly depends on its genre-characteristics. This is one of the key issues of this project (see our pilot study in Berzlánovich et al. 2008).

In our view three subsystems build up the overall cohesion of a text: referential cohesion, relational cohesion and lexical cohesion. The elements of referential cohesion are cohesive by sharing identity of reference. Relational cohesion is created by signaling the relations between EDUs in the text with connectives or ellipsis. Lexical cohesion looks at the semantic network of the lexical elements in the text. In short, by definition all these three types of cohesion reveal semantic relations of the surface items in the text. In our project we look at the relations between content words (more about the selected items in section 3.3). Table 1 shows our categorization for lexical cohesive relations.

Table 1: Categories of lexical cohesion

Category		Example
Repetition		<i>sun – sun</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>
Non-systematic semantic relations	Collocation	<i>light - star</i>

We distinguish three main types of lexical cohesion: repetition, systematic semantic relations, and non-systematic semantic relations. By *repetition* we mean word repetition. Under the heading *systematic semantic relations* we include the traditional semantic relations that are known from lexical semantics: hyponymy, hyperonymy, co-hyponymy, meronymy, holonymy, co-meronymy, synonymy and antonymy. The label *non-systematic semantic relations* stands for the collocation relations. Problems related to these categories are discussed in the rest of this section.

Firstly, general problems are considered that are relevant to all the relations. The first question concerns the relevance of context in the analysis of the lexical relations. Studies (McCarthy 1988, Tanskanen 2006) with a discourse-specific approach as opposed to the lexical-semantic approach analyze lexical cohesive relations in their context. Their aim is to focus on the “communicative potential” rather than on the “meaning potential” of the items which is the case in the lexical-semantic approach. We take the lexical meaning when identifying the relations, and ignore those “instantial relations” (an example from Tanskanen 2006, p. 67: *the Arabs – oil*) that arise from the context. Still, it is not always easy to decide which meaning relation is context-free. Take the example of *Mercury – axis/pole/crater*. It is hard to decide how identifiable these relations are without context. The reader with general world knowledge about the shape of the planets (for example, the Earth), may perceive the relations of the item *Mercury* to *axis* and to *pole* without context. However, further background knowledge or textual context may be necessary to establish the relation between *Mercury* and *crater*. In most cases, systematic semantic relations are easy to identify without context in contrast to collocations, which are often identified in their context. This is related to the question of register-sensitive and domain-sensitive relations. We identify text-general relations; i.e., relations which are not specific of a certain register or domain.

When annotating the relations, we keep track of whether the items in the relation have identity of reference, but we do not divide the relations with identity of reference and non-identity of reference into separate (sub)categories (unlike e.g., Halliday & Matthiessen 2004). We argue that the identity of reference does not influence whether a semantic relation does or does not exist in the text. (However, it may play a role when making decisions about the strength of the relations – see section 3.4.)

Finally, when we encounter problematic cases of any lexical cohesive relation, we use the Cornetto database (Vossen, Hoffmann, de Rijke, Tjong Kim Sang & Deschacht 2007). The Cornetto database is a semantic database for Dutch. As it integrates the Dutch WordNet (DWN) with another semantic resource for Dutch (Referentie Bestand Nederlands, RBN), it combines vertical semantic relations and horizontal relations. It is thus a reliable source in case of doubt for both the systematic and non-systematic semantic relations.

After discussing the general problems when identifying the relations, we now discuss the specific problems for each category. *Repetition* is the reoccurrence of words in the text. Halliday and Hasan (1976) and Tanskanen (2006) keep repetition together with systematic semantic relations under one main

category (reiteration). In this paper repetition is kept separately from the systematic and non-systematic semantic relations for two reasons. Firstly, the meaning of this relation is “identification of concepts”. The cohesive force between the lexical items arises by saying that the same concept is discussed, and by the lexical items being realized by the identical lemma. Secondly, the repetition of a lemma can be easily followed in a transitive chain in the text. Building a transitive chain of other lexical cohesive relations is an uncommon phenomenon (as shown in section 3.1).

*Hyponymy* and *hyperonymy* are lexical cohesive relations between an item and a more general item. In the case of hyperonymy the general item creates a cohesive link with the preceding more specific item, whereas in the case of hyponymy the more specific item creates the link with the preceding general item. Thus the directionality of the relation makes the difference between them.

*Meronymy* and *holonymy* are lexical cohesive relations between two items – one item being part or member of the other item. For meronymy, the item expressing ‘part’ or ‘member’ builds a cohesive link with the first item expressing ‘whole’. Conversely, for holonymy the ‘whole’ item creates a cohesive relation linking back to the ‘part’ or ‘member’ item. Similarly to the distinction between hyponymy and hyperonymy, the directionality of the relation determines under which category the relation is ranked.

The relations *hyponymy*, *hyperonymy*, *meronymy* and *holonymy* are formally transitive (unlike synonymy and antonymy). However, relatedness may drop dramatically for hierarchies above two levels. The question is when item *a* is related to item *b*, and *b* related to *c*, whether *a* is related to *c*. In other words, how far can we go when looking at the general items (for hyponymy and hyperonymy) and the whole item (for meronymy and holonymy)? For example, if *handle* is part of the *door*, and *door* is part of the *house*, is there a meronymic relation between *handle* and *house* (Stokes 2004)?

The *co-hyponymy* relation links two specific items which share a common general item. In other words, they are two kinds or instances of the general item. Along similar lines, *co-meronymy* links two items which are parts or members of the same item. Tanskanen (2006) merges co-hyponymy and co-meronymy into one category. In fact, there were instances in our analysis for the pilot study (Berzlánovich et al. 2008) where it was hard to decide which category label to apply to a relation. It was especially problematic in our subcorpus of encyclopedia entries on astronomy, where general items (*planet*, *star*) and more specific items (*dwarfstar*, names of stars (*sun*, *Proxima Centauri*) and names of planets (*Mercury*, *Venus*, *Earth*)) occurred in a large number. When looking at the examples in Table 1, the question arises why *Venus* and *Mercury* are co-hyponyms, but *Earth* and *sun* co-meronyms. All these stellar objects are the members of the solar system, thus all could be taken as co-meronyms of each other. We made a distinction between the celestial objects according to which category they belong to in the solar system: *sun* as a star; *Mercury*, *Venus* and *Earth* as instances of the eight planets. Hence, *Venus* and *Mercury*, both being planets are co-hyponyms. However, *Earth* as a planet and *sun* as a star are members of the solar system; these items are thus co-meronyms.

Consider another problematic case. For the lexical items *minute* and *kilometer* we identified the relation co-hyponymy. One reason for this decision is that the WordNet database (the source for dubious cases in the pilot study) ranks both items under the abstract category of ‘measure’. Another reason is that both for *minute* and *kilometer* the WordNet database provided the definition “measuring distance”.

These difficulties lead to a more general question: What is the highest degree of the abstraction that still should be included in the analysis? This question applies to the relations hyponymy, hyperonymy, meronymy and holonymy, where the general item is overt, and also to co-hyponymy and co-meronymy, where the general item remains covert. For the former (vertical) relations, items that are too distant in the hierarchy should not be considered related; for the latter (horizontal) relations, multiple available superordinates may yield different relations (as we have seen above). A possible solution for such cases might be to assign more than one category label to the link between the lexical items. However, this strategy may lead to methodological difficulties both in the quantitative and qualitative analysis. Our decision is to distinguish *co-hyponymy* and *co-meronymy* unlike Tanskanen (2006), and to assign one label to each lexical cohesive link. Similarly to our method in the pilot study (using WordNet for unclear cases) we use the Cornetto database to distinguish co-meronymy from co-hyponymy and to identify them in case of doubt in future analysis.

*Synonymy* is a broad category in Halliday (1985) where hyponymy and meronymy are variants of synonymy. We follow the categorizations (e.g., Halliday & Hasan 1976, Hasan 1984, Martin 1992, Tanskanen 2006) where synonymy is defined in a narrower sense. Synonymy is a relation between lexical elements whose sense is the same or nearly the same. We had the same definition for *repetition*. The difference between these two categories is the identity of lemmas in the case of repetition as opposed to the non-identity of lemmas in synonymic relations. It is hard to find clear synonyms, as usually the items occur in different contexts or in different registers, or there are a few traits that differ between the lexemes. They cannot thus always be replaced with each other (Cruse 1986).

*Antonymy* relates two items with opposite senses. This again is not a clear-cut category. Note Halliday and Hasan’s (1976, p. 285) examples for pairs of opposites of various kinds: “complementaries such a *boy ... girl, stand up ... sit down*, antonyms such as *like ... hate, wet ... dry, crowded ... deserted*, and converses such as *order ... obey*”. For a more detailed discussion on antonyms see Cruse (1986). The phenomenon of blurred boundaries between senses of lexemes in the case of both synonymy and antonymy might have been the motivation for Halliday and Matthiessen (2004) to rank both antonymic and synonymic relations under the category of synonymy.

The third type of lexical cohesion is *collocation*. It has to be emphasized that the term *collocation* for lexical cohesive relations differs from what is meant by collocation in corpus linguistics. In corpus linguistics *collocation* refers to co-occurrences of words in the particular texts. Many of these words co-occur in a

fixed syntactic pattern (e.g., *make an improvement; a high/enormous/greater/large/mild/reasonable/substantial degree of...*) (Stubbs 2001). *Collocation* in lexical cohesion exists between words in similar textual context. The difference between the two interpretations of the term is that in lexical cohesion the mere co-occurrence of items is not a sufficient criterion. We analyze lexical items with a meaning relation between them, 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 & Hirst 1991, p. 22). This definition is still vague, and it makes collocation probably the most problematic category for the analysis of lexical cohesion. Hasan (1984) does not include it into her categorization as a separate category, but incorporates some cases of collocation into other lexical categories. She argues that in the case of collocation “the problems of inter-subjective reliability cannot be ignored” (Hasan 1984, p. 195). However, systematic semantic relations might be inter-subjective as well (see the examples discussed above, and Morris, Beghtol & Hirst 2003). Collocation is an often neglected category of lexical cohesion in computational linguistics as well, as it is more difficult to detect automatically (Morris & Hirst 1991). An approach to identify collocations automatically is based on word co-occurrence statistics (Stokes 2004). Although it captures a number of collocation relations, it still cannot be regarded sufficient for the automatic identification of all relevant cases.

The identification of collocation relations is not easy indeed. It has been argued that collocation is context-dependent (Morris & Hirst 2004), which makes the identification even more difficult. Note, however, collocations (e.g., Morris et al. 2003, p. 62: *broom – sweep*) where no context is necessary for their identification – similarly to systematic semantic relations. Furthermore, knowledge structures (frames, scripts and schemas) help the identification of certain collocation relations (Tanskanen 2006). With this, collocation is “located within the system of the world, not the system of language” (Morris et al. 2003, p. 160). Patterns of collocation can definitely be revealed for certain cases. Halliday and Matthiessen (2004, p. 577) identify “circumstantial relationships” (*dine – restaurant, bake – oven*) and “participant + process relationships” (*smoke – pipe*). The latter involves further subcategories: “Process + Range (e.g., *play + musical instrument: piano, violin, etc.; grow + old*) and Process + Medium (e.g. *shell + peas, twinkle + star, polish + shoes*)”. Martin’s (1992) nuclear relations and Tanskanen’s (2006) activity-based relations refer to similar examples. For some of the examples case relations can be defined (e.g., instrumental case for *smoke – pipe*), but not all the collocations can be identified in this way (Morris et al. 2003). We rely on these observations from previous studies, and use our intuitions on frame, script and schema structures in our analysis. In case of uncertainty, the Cornetto database is used in the following way: If the definition of a lexeme contains the other lexical item in question, we take these two elements as forming a collocation relation.

### 3.3 Elements forming lexical cohesive relations

In this section we focus on the items that are selected for the investigation of lexical cohesion. The major questions concern the part of speech of the items and multi-word units.

The question of which parts of speech to include in the analysis, is highly relevant in lexical cohesion. Certain relations tend to be realized with items of a certain part of speech. For instance, while collocation relations often contain verbs, hyponymy/hyperonymy and meronymy/holonymy are typical meaning relations between nouns. The early work of Halliday and Hasan (1976) makes a clear distinction between grammatical items (members of a closed system, for example, personal pronouns and demonstratives) and lexical items (members of an open set) according to their contribution to cohesion. The former participate in referential cohesion, substitution and ellipsis, whereas the latter in lexical cohesion. Hasan (1984) proposes to eliminate this strong division by looking at the interaction of referential chains and similarity chains arising from systematic lexical cohesive links. As regards referential chaining, lexical items are included in the analysis besides pronouns, demonstratives and the definite article in certain investigations, while they are excluded in others. As for lexical cohesion, the biggest differences concern the part of speech of the content words used for the analysis.

Many computational studies (e.g., Barzilay & Elhadad 1999, Silber & McCoy 2002, Stokes 2004) select exclusively nouns for analysis for practical reasons. Their main argument is that verbs (being more polysemous than nouns, and few of them being “truly synonymous”) are more difficult to be investigated automatically. Besides, they are difficult to be ranked under certain categories of lexical cohesion. For this reason, when designing the WordNet database, many verbs were ranked under the category *entailment* which is traditionally not a lexical cohesive category. For instance, *snore* entails *sleep*, “as a person cannot snore unless he/she is sleeping” (Stokes 2004, pp. 12-13).

So far nouns have been in the focus of the investigations into lexical cohesion. Our aim is to widen this perspective. In our project we look at nouns, verbs, adjectives and certain instances of adverbs as candidate items for lexical cohesive relations. This decision is motivated by our other decision that for the identification of cohesive relations we take lemmas instead of word forms as building elements of a cohesive link.

In our analyses we have ignored word class and inflectional and derivational differences so far. No distinction has been made between the relations according to these differences, but this decision may have to be reconsidered in the course of our project. It might have a role when making decisions about the strength of the relation. For instance, repetition of identical word forms might be stronger than the repetition where the items differ in derivation.

Moreover, differences in the word forms lead to difficulties which lexical cohesive relation to choose. Previous studies unanimously treat different inflectional and derivational word forms with an identical lemma as repetition links. We agree with the decision to treat inflectional variants as repetitions, as the inflectional affix only slightly modifies the meaning (e.g., *high* – *highest*, *type* –

*types*). However, derivational word forms need further consideration in our project, as derivational affixes may dramatically change the meaning of the word. So far we have treated different derivational word forms as repetitions (e.g., *ster* ‘star’ – *sterretje* ‘little star’). Similarly, *Earth* and *earthly* form a repetition link. Since we thus ignored derivational differences, *earthly* and *Mercury* were regarded as co-meronyms in our pilot project (Berzlánovich et al. 2008). Halliday and Matthiessen (2004, p. 572) follow the same approach, but they note that it is hard to decide in certain cases of derivational variants (e.g., *rational* – *rationalize* – *ration* – *reason*) if they are “close enough to be considered the same item”. They argue that derivational variants based on a still productive derivational process can be regarded as identical. This might suggest that a distinction could be made between derivational forms on the basis of the productivity of the affix. Take, however, the example of *write* and *writing*<sub>1</sub> (writing as a process) and *writing*<sub>2</sub> (writing as the product of the process). While for *write* and *writing*<sub>1</sub> a repetition link may easily be assigned, *write* and *writing*<sub>2</sub> fitting into our frame concept may be ranked under collocation. In order to make a final decision on the status of derivational forms, an inventory is being built from the corpus in our project.

We have to set certain constraints for the selected word classes. As for the nouns, both common nouns and proper nouns are included in the analysis. Proper names seem especially relevant in our subcorpus of encyclopedia entries on astronomy. Here meaning relations are formed between stellar objects which are identified with their names (*Mercury*, *Proxima Centauri* etc.). In the case of verbs we ignore *have* and *be* as main verbs due to their very general meaning. Modal verbs and auxiliaries are always regarded as parts of the verb phrase, i.e., word forms of the given lemma. We make restrictions for adverbs as well. Traditional grammars define the word class of adverbs in different ways. As our corpus consists of Dutch texts, we follow the Dutch grammar book *Algemene Nederlandse Spraakkunst (ans)*. From the meaning categories of the adverb we take the place adverbs, time adverbs and the adverbs of frequency as candidates for participating in lexical cohesion. The other categories of adverbs defined in the *ans* (adverbs of graduality, adverb intensifiers, adverbial quantifiers, modality adverbs, the adverb of negation *niet*, conjunctive adverbs, pronominal adverbs, the adverb *er* and prepositional adverbs) are closer to function words (particles, conjunctions and prepositions) than to the open class of content words. It might be argued that function words also bear meaning; therefore, they should be included in the analysis. However, similarly to the main verbs *have* and *be* they have a general meaning. Moreover, in most cases they would form repetition relations. It does not seem plausible that the repetition of function words (other than those with discourse structuring function) would contribute to the cohesion of the text.

Numerals are neglected in most cohesion studies as well. An interesting counterexample is Taboada (2004), where numerals were found to have great cohesive force in the analyzed corpus (spoken task-oriented dialogues). This draws attention again to the genre differences concerning lexical cohesion. We excluded numerals from the analysis. One reason is that it would be hard to define a relation between numerals, except of repetition. Generally, numerals do not tend

to form semantic relations with other items. A typical use of numerals is the ‘numeral + measure’ (e.g., € 2.50) in our texts. A few meaning relations can be established between such phrases from the context; for example, € 2.50 – *small amount*. However, we are interested in text-general, context-free relations. A context-free relation is € 2.50 – *amount*. Still, we excluded this relation from our analysis. The reason for this decision is that we take single words, but not multi-word items as the building elements of the lexical cohesive relations. Hence, we took simply € – *amount* as a lexical cohesive relation (collocation).

The issue of multi-word items is related to decomposability (Copestake et al. 2002). We take the *lexical item* as the unit for the cohesion analysis. This is the smallest lexical unit whose meaning cannot be decomposed. We are aware of the fact that there are semantically indecomposable *multi-word units* (for example, idioms). The few instances of idioms found in our corpus so far, however, do not participate in lexical cohesion. So far we have encountered only one multi-word unit entering a lexical cohesive relation which cannot be decomposed semantically. This single example (*Proxima Centauri*) is a proper name, and its components do not exist as independent lexical items. We also consider *compounds* as indecomposable lexical items, and do not look at the components separately. We argue that the meaning of a compound is always more than the sum of the meaning of its components. With this we do not deny that the construction of the compounds is motivated by the meaning of their components, and the structure of many compounds is more transparent (e.g., *high + land* → *highland*) or less transparent (e.g., *high + way* → *highway*). In spite of such differences we treat all compounds in the same way in order to maintain systematicity in our analysis. Hence, for example, the compounds *dwarfstar* – *dwarf phase* do not create a repetition link (although their first components are identical), but looking at the meaning of the compounds as a whole we take the lexical cohesive relation between them as collocation. Eventually, our criteria for the selection of the items are semantic. The identity of form is relevant exclusively for the distinction between repetition and synonymy. For this reason, instances of homonymy (e.g., *light*<sub>adjective</sub> – *light*<sub>noun</sub>) are not included in our analysis either.

### 3.4 The strength of lexical cohesive relations

Discourse is hierarchically built. The question concerning lexical cohesion is whether and how this structure can be captured. The strength of the lexical cohesive relations depends on the distance of the two items creating the relation and the type of the relation they form. Computational approaches often combine these two parameters when measuring the strength of the relations (e.g., Silber & McCoy 2002). Besides these two criteria, other factors influencing the strength of lexical cohesive chains are the number of repetitions, the length of the chain, the distribution of the cohesive items in the text, and chain density (Barzilay & Elhadad 1999). Since we analyze lexical cohesion as a network of relations, we focus on the parameters of distance and the type of relation between the two lexical items.

The distance of the items covers two aspects. Firstly, the physical distance of two lexical items in the linear text determines how salient the relation is in the text. In general, the longer the distance between two items, the weaker the relation. The second aspect of distance means the semantic distance between the items. The closer the semantic distance, the stronger the relation is between the two items. Semantic distance has been measured by the path length in semantic databases in many cases (Budanitsky and Hirst 2006). The common problem with the use of semantic networks is that the vertical distance between the items cannot be measured by the number of the semantic levels, as these are not evenly distant from each other. The WordNet database, for example, is structured in a way that vertical distance is larger at higher levels in the hierarchy. Another problem concerning the use of databases is that collocation relations cannot be measured with them. Furthermore, the databases often do not contain all the lexical items of the analyzed corpora – even though new entries are added to some databases (e.g., WordNet) (Stokes 2004).

The final problem concerning the use of lexical semantic databases is the question of genre and domain. Certain databases (for example, WordNet) are built up of general relations not depending on a special domain. They might thus be insufficient for the analysis of corpora within a certain domain (for example, imix – a medical corpus in Dutch). The main corpus of the Cornetto database used for our project is primarily a global database, and it contains also a sublexicon for the domain of financial law. We expect to encounter relations in our corpus that are not included in the main corpus in the Cornetto.

The strength of the lexical cohesive relations also depends on the type of the relations. Repetition is commonly considered as the strongest relation (e.g., Stokes 2004). However, ranking the systematic semantic relations and collocation according to their strength is problematic due to the unclear status of collocation.

Two more factors that are usually ignored in the previous studies concern the identity of reference and the form of the words entering a lexical cohesive relation. It is expected that relations with items of identity of reference are stronger than relations without identity of reference – regardless of the type of the relation. However, the identity of reference can be investigated only in the case of nouns, but not for verbs, adjectives or adverbs. Concerning word forms, it might be assumed that relations where the lexical items have identical word form are stronger than relations where items differ in word class or in their derivational forms.

For the investigation of the structuring of lexical cohesion we aim to devise a measurement. In our pilot study (Berzlánovich et al. 2008) the central discourse units were measured as follows. For the identification of the most central EDU we took the EDU which created the most lexical cohesive relations with other EDUs. The most central larger discourse units were linked with the most lexical cohesive relations to other discourse units.

However, taking simply the number of the relations into account is insufficient for measuring lexical cohesion. The development of a more precise method is future work in the project. We will include the number of lexical cohesive relations as it proved to be a good, though not sufficient indicator to

measure the centrality in lexical cohesion. It has still not been decided whether we should count the relations per lexical item, per EDU and/or per larger discourse unit. Then the textual distance between two lexical items has to be considered. We can measure the number of the intervening EDUs or lexical items. As the length of the particular EDUs differs, counting the intervening lexical items might be a more reliable indicator. Ranking the types of lexical cohesive relations might also contribute to measuring the strength of the relations. As the identity of form is an additional criterion to the semantic one in the case of repetition, we may agree with the decision of previous studies (e.g., Silber & McCoy 2002, Stokes 2004), and take it as the strongest relation. It is a question how to measure semantic distance for the other relation types. Relying on the Cornetto database to measure semantic distance may lead to difficulties, as earlier studies have already pointed to the problems when using WordNet for this purpose. The controversial claims about the status and importance of collocation in lexical cohesion make it difficult to decide whether this type of relation should be considered weaker, stronger or equal compared to systematic semantic relations. Finally, the factors of the identity of reference and the identity of word form may be excluded from our measures, as they are not applicable to all lexical items and relation types.

### 3.5 Conclusion

The *pattern of lexical cohesion* is a network which results from the multiple relations of the lexical items. In any type of lexical cohesive relation we look at the closest occurrence of the preceding element that forms a relation with the second element. Due to their transitivity repetition relations form a chain in contrast to other types of lexical cohesive relations.

The three main *types of cohesion* are referential, relational and lexical cohesion. The three main categories of lexical cohesion are repetition, systematic semantic relations (for example, hyponymy, meronymy and synonymy) and non-systematic semantic relations (i.e., collocations). We identify relations arising from lexical meaning, and ignore “instantial” meaning relations arising from context. We use the Cornetto database for the identification of lexical cohesive relations in dubious cases.

*Content words* (nouns, verbs, adjectives, place and time adverbs and adverbs of frequency) are candidate items for participating in lexical cohesion. Proper names are included in the lexical cohesion analysis. The elements of multi-word units (except proper names) are analyzed as separate items entering a lexical cohesive relation. We take compounds and idioms as indecomposable single units.

The *strength of the lexical cohesive relations* depends on the number of the relations, the textual and semantic distance and the type of relation between the two items forming the link. Based on these factors a measurement will be devised for lexical cohesion.

## 4 Outline of the project

### 4.1 Research questions

The project investigates the hierarchical organization of discourse. We examine discourse organization by looking at the coherence structure and the lexical cohesion of different genres. The main research question is the following:

**What is the interaction between cohesion and coherence in expository and persuasive genres?**

This question is investigated in two main steps. The hierarchical discourse organization is examined first at the *global levels*. As both coherence and lexical cohesion vary with genre, the differences for expository and persuasive texts are investigated.

RQ1: Which lexical cohesive relations characterize the expository and/or persuasive texts?

RQ2: Which coherence relations characterize the expository and/or persuasive genres?

The coherence structure shows the discourse units of the texts. Its hierarchical structure reflects the more and less central discourse units. The most central discourse units are identified in the lexical cohesion analysis as well. The next research question concerns the alignment between these two structures. We expect genre differences in the alignment for various genres. Our method is to map the move structure of a genre onto the coherence structure of the particular texts, and compare it with the lexical cohesion of each unit. The degree of the alignment is captured with comparing the central and marginal units identified with the coherence structure and lexical cohesion.

RQ3: How close is the alignment between the coherence structure and lexical cohesion in expository and persuasive genres at the global level?

After getting a global view of the structure of genre, coherence and cohesion, the interaction between coherence and lexical cohesion is examined *at lower levels*.

RQ4: How are lexical cohesive relations distributed within and between the larger discourse units in expository and persuasive genres?

RQ5: How are coherence relations distributed within and between the larger discourse units in expository and persuasive genres?

We expect to find systematic connection between certain coherence relations and lexical cohesive relations. It is an open question whether genre differences concerning the alignment can be found at local levels as well.

RQ6: How close is the alignment between the coherence structure and lexical cohesion in expository and persuasive genres at lower levels in the hierarchy?

## 4.2 Corpus

For the investigation of the research questions we build a corpus comprising expository and persuasive genres. We collect 50 expository texts and 50 persuasive texts. The general criteria for the selection of the texts are that they are well-edited by professional writers, they are written in Dutch, and they have a minimum length of 250 words. For expository texts 25 encyclopedia entries and 25 popular scientific articles will be included. For persuasive texts 25 fundraising letters and 25 advertisements will be collected. These genres can be ranked on a scale being strongly expository (encyclopedia entries) at one end, and strongly persuasive (advertisements) at the other end. We also add 25 book reviews and 25 film reviews to the corpus. Reviews are so-called mixed genres containing both expository and persuasive text parts. They can thus be placed somewhere between expository and persuasive texts on the scale. The text parts describing e.g. the story line or the characters are expected to bear similarities with the structure of the expository texts, whereas the evaluative text parts are expected to have similar structure to persuasive texts.

## 4.3 Methods

For defining the **genre structure**, we follow the move analysis (see 2.2.2). In the case of fundraising letters we follow the moves already defined in previous research by Upton (2002). For advertisements, we modify the move structure suggested by Bhatia (2005). For genres where there is no literature on move structure available, we build up the move structure on the basis of generalizations about the genre in previous literature, and our observations when analyzing the corpus. For instance, in the case of the expository encyclopedia entries the labels for the moves (*name*, *define*, *describe in general*, *describe details*) have been inductively derived from the corpus. These labels are supported by previous findings on the generic structure of expository texts (Britton 1994, Berman & Nir-Sagiv 2007).

For the analysis of **coherence**, we follow Rhetorical Structure Theory (RST) (Mann & Thompson 1988, Taboada & Mann 2006). In this tradition the functional relations between the discourse units are defined from the writer's perspective (based on the analysts' plausibility judgements and semantic criteria in the relation definitions). The coherence structure of a text combines subject matter relations (relating states of affairs) and presentational relations (relating illocutions or text parts) in a single representation.

For the analysis of **lexical cohesion** we identify lexical semantic relations between content words. We distinguish the following lexical cohesive relations: repetition, systematic semantic relations and non-systematic semantic relations (collocation) (for details see section 3).

Interrater reliability tests will be carried out for the analyses of moves, coherence and lexical cohesion. This will be done for part of the corpus.

The **hierarchy** of discourse is investigated both in the coherence structure and in lexical cohesion. Following the RST tradition we distinguish more and less central EDUs (*nuclei* and *satellites*) in the coherence structure. The coherence relations between the discourse units apply recursively to yield a hierarchical structure. To build up a hierarchy in lexical cohesion, a measurement or a set of measurements will be developed in the project. The factors we need to take into consideration are: the number of the lexical cohesive relations, the textual and semantic distance between the lexical items and the type of relation. It is still an open question whether the identity of reference and differences in the form of the related items should be included (see section 3.4).

## References

- Bärenfänger, M., Lobin, H., Lungen, H., & Hilbert, M. (2006). Using OWL ontologies in discourse parsing. In K.-U. Kühnberger & U. Mönnich (Eds.), *Proceedings of the Workshop of Ontologies in Text Technology*. Osnabrück, Germany, September 28-29, 2006.
- Barzilay, R. & Elhadad, M. (1999). Using lexical chains for text summarization. In I. Mani & M. T. Maybury (Eds.), *Advances in automatic text summarization*. (pp. 111-121). Cambridge, MA: MIT Press.
- Beaugrande, de R., & Dressler, W. (1981). *Introduction to text linguistics*. London: Longman.
- Beigman Klebanov, B. & Shamir, E. (2006). Reader-based exploration of lexical cohesion. *Language Resources and Evaluation* 40, 109-126.
- Berman, R. A. & Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: a developmental paradox. *Discourse Processes* 43(2), 79-120.
- Berzlánovich, I., Egg, M. & Redeker, G. (2008). Coherence structure and lexical cohesion in expository and persuasive texts. In A. Benz, P. Kühnlein & M. Stede (Eds.), *Proceedings of the Workshop Constraints in Discourse III* (pp. 19-26). Potsdam, Germany, July 30 – August 1, 2008.
- Bestgen, Y. (1998). Segmentation markers as trace and signal of discourse structure. *Journal of Pragmatics* 29(6), 753-763.
- Bestgen, Y. & Vonk, W. (2000). Temporal adverbials as segmentation markers in discourse comprehension. *Journal of Memory and Language* 42(1), 74-87.
- Bhatia, K. V. (2005). Generic patterns in promotional discourse. In: H. Halmari & T. Virtanen (Eds.), *Persuasion across genres* (pp. 213-225). Amsterdam: Benjamins.
- Bolshakov, I. & Gelbukh, A. (2001). Text segmentation into paragraph based on local text cohesion. In V. Matousek, P. Mautner, R. Moucek & K. Tauser (Eds.), *Text, speech and dialogue* (pp. 158-166). 4<sup>th</sup> International Conference, TSD 2001; Železná Ruda, Czech Republic, September 11-13, 2001. Berlin: Springer.
- Britton, K. B. (1994). Understanding expository text. Building mental structures to induce insights. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 641-674). San Diego: Academic Press.
- Budanitsky, A. & Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics* 32(1), 13-47.
- Buitkienė, J. (2005). Variability of cohesive devices across registers. *Studies About Languages (Kalbų Studijos)* 7, 17-20.
- Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics* 22, 75-95.
- Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag A. I. & Flickinger, D. (2002). Multiword expressions: linguistic precision and reusability. In M. González Rodríguez & C. P. Suárez Araujo (Eds.), *Proceedings of the 3rd International Conference on Language Resources*

- and *Evaluation (LREC 2002)*, (pp. 1941-1947). Las Palmas, Canary Islands, 2002. Paris: ELRA.
- Cruse, D. A. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.
- Enkvist, N. E. (1990). Seven problems in the study of coherence and interpretability. In U. Connor & A. M. Johns (Eds.), *Coherence in writing: Research and pedagogical perspectives* (pp. 9-28). Washington, DC: TESOL.
- Fellbaum, Ch. (Ed.), (1998). *Wordnet. An electronic lexical database*. Cambridge, Mass: MIT Press.
- Ferret, O. (2002). Using collocations for topic segmentation and link detection. In S.-C. Tseng, T.-E. Chen & Y.-F. Liu (Eds.), *Proceedings of the 19<sup>th</sup> International Conference on Computational Linguistics*. Taipei, Taiwan, August 24-September 1, 2002.
- Ferret, O. (2007). Finding document topics for improving topic segmentation. In *Proceedings of the 45<sup>th</sup> Annual Meeting of the Association of Computational Linguistics* (pp. 480-487). Prague, Czech Republic, June 23-30, 2007.
- Filippova, K. & Strube, M. (2006). Using linguistically motivated features for paragraph boundary identification. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 06)* (pp. 267-274). Sydney, Australia, July 22-23, 2006.
- Fox, B.A. (1987). *Discourse structure and anaphora: Written and conversational English*. Cambridge: Cambridge University Press.
- Geerts, G., Haeseryn, W., de Rooij, J. & van den Toorn, M. C. (Eds.), (1984). *Algemene Nederlandse spraakkunst*. Groningen: Wolters-Noordhoff.
- Goutsos, D. (1997). *Modelling discourse topic: sequential relations and strategies in expository text*. Norwood, NJ: Ablex.
- Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Arnold.
- Halliday, M. A. K. & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3<sup>rd</sup> ed.). London: Arnold.
- Hasan, R. (1984). Coherence and cohesive harmony. In J. Flood (Ed.), *Understanding reading comprehension: Cognition, language and the structure of prose* (pp. 181-219). Newark, DE: International Reading Association.
- Hellman, Ch. (1995). The notion of coherence in discourse. In G. Rickheit & Ch. Habel (Eds.), *Focus in coherence in discourse processing* (pp. 190-202). Berlin: Gruyter.
- Hirst, G. & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *Wordnet. An electronic lexical database* (pp. 305-332). Cambridge, MA: MIT Press.
- Hodgson, J. M. (1991). Informational constraints on pre-lexical priming. *Language and Cognitive Processes* 6(3), 169- 205.
- Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

- Hoey, M. (2005). *Lexical priming: a new theory of words and language*. London: Routledge.
- Korfiatis, G. (2007). *Discourse structure in Dutch: Semi-automatic annotation*. First year report. Center for Language and Cognition Groningen, University of Groningen.
- Louwerse, M. M., McCarthy, Ph. M., McNamara, D.S. & Graesser, A.C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the 26<sup>th</sup> Annual Meeting of the Cognitive Science Society* (pp. 843-848). Mahwah, NJ: Erlbaum.
- Mann, W. C. & Thompson, S. A. (1988). Rhetorical structure theory: toward a functional theory of text organization. *Text* 8(3), 243-281.
- Martin, J. R. (1992). *English text. System and structure*. Amsterdam: Benjamins.
- McCarthy, M. (1988). Some vocabulary patterns in conversation. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching*. London: Longman.
- Morris, J., Beghtol, C. & Hirst, G. (2003). Term relationships and their contribution to text semantics and information literacy through lexical cohesion. In *Proceedings of the 31<sup>st</sup> Annual Conference of the Canadian Association for Information Science* (pp. 153-168). Halifax, Nova Scotia, June 1-June 4, 2003.
- Morris, J. & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17(1), 21-48.
- Morris, J. & Hirst, G. (2004). Non-classical lexical semantic relations. In D. Moldovan & R. Girju (Eds.), *Proceedings of the Workshop on Computational Lexical Semantics* (pp. 46-51). Boston, MA, USA, May 6, 2004.
- Polanyi, L., Culy, Ch., van den Berg, M., Thione, G. L. & Ahn, D. (2004). A rule based approach to discourse parsing. In M. Strube & C. Sidner (Eds.), *Proceedings of the 5<sup>th</sup> SIGdial Workshop on Discourse and Dialogue* (pp. 108-117). Cambridge, MA, USA, April 30-May 1, 2004.
- Redeker, G. (2000). Coherence and structure in text and discourse. In H. Bunt & W. Black (Eds.), *Abduction, belief and context in dialogue. Studies in computational pragmatics* (pp. 233-263). Amsterdam: Benjamins.
- Sanders, T. J. M. & Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes* 29(1), 37-60.
- Silber, H. G. & McCoy, K. F. (2002). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics* 28(4), 487-496.
- Stokes, N. (2004). *Applications of lexical cohesion analysis in the topic detection and tracking domain*. Ph.D. dissertation, Department of Computer Science, University College Dublin, Dublin.
- Stubbs, M. (2001). Computer-assisted text and corpus analysis: lexical cohesion and communicative competence. In D. Schiffrin, D. Tannen & H. E.

- Hamilton (Eds.), *The handbook of discourse analysis* (pp. 304-320). Blackwell.
- Swales, J. (1990). *Genre analysis. English in academic and research settings*. Cambridge: Cambridge University Press.
- Taboada, M. T. (2004). *Building coherence and cohesion*. Amsterdam: Benjamins.
- Taboada, M. & Mann, W. C. (2006). Rhetorical structure theory: looking back and moving ahead. *Discourse Studies* 8(3), 423-459.
- Tanskanen, S.-K. (2006). *Collaborating towards discourse: lexical cohesion in English discourse*. Amsterdam: Benjamins.
- Thompson, S. (1994). Aspects of cohesion in monologue. *Applied Linguistics* 15(1), 58-75.
- Upton, T. A. & Connor, U. (2001). Using computerized corpus analysis to investigate the text linguistic discourse moves of a genre. *English for Specific Purposes* 20, 313-329.
- Verikaitė, D. (2005). Variation of conjunctive discourse markers across different genres. *Man and the Word (Žmogus ir žodis)* 3(7), 68-75.
- Vossen, P., Hofmann, K., de Rijke, M., Tjong Kim Sang, E. & Deschacht, K. (2007). The Cornetto Database: Architecture and User-Scenarios. In M.-F. Moens, T. Tuytelaars & A. P. de Vries (Eds.), *Proceedings of the 7<sup>th</sup> Dutch-Belgian Information Retrieval Workshop* (pp. 89-96). DIR 2007. Leuven, Belgium, March 28-29, 2007. Leuven: ACCO Press.
- Yankova, D. (2006). Semantic relations in statutory texts: A study of English and Bulgarian. *SKY Journal of Linguistics* 19, 189-222.

## Appendix A: First year activities

### 1 Research activities

Aug 2007-Oct 2007	Reading on coherence
Aug 2007-	RST-training
Nov 2007-	Preparatory explorations for corpus collection (NWO Programme team)
Nov-Dec 2007	Reading on corpus linguistics
Jan-Feb 2008	Reading on cohesion
Feb 2008-	RST-analyses (NWO Programme team)
Feb 2008	Developing the methodology of cohesion analysis
Feb 2008-	Cohesion analyses
March-May 2008	Reading on genre theory and text typology
May-June 2008	Preparing conference presentations
June 2008	Reading on discourse structure
June-July 2008	Writing proceedings paper for <i>Constraints in Discourse III</i> ; preparing presentation

### 2 Output

#### 2.1 Conference presentations

- Berzlánovich, I. (2008). Genre-specific organization of coherence and cohesion in discourse. *TABU Day*, Groningen, 6 June 2008.
- Berzlánovich, I., Egg, M. & Redeker, G. (2008). Coherence structure and lexical cohesion in expository and persuasive texts. *Constraints in Discourse III*, Potsdam, 30 July-1 August 2008.

#### 2.2 Publication

- Berzlánovich, I., Egg, M. & Redeker, G. (2008). Coherence structure and lexical cohesion in expository and persuasive texts. *Proceedings of the Workshop Constraints in Discourse III*, Potsdam, 30 July-1 August 2008.

### 3 Courses

#### 3.1 University of Groningen

Autumn 2007-2008	Corpus linguistics (van Noord, G. & van der Plas, L.) Discourse and pragmatics (Egg, M., Mazeland, H. & Redeker, G.) Linguistic analysis (Hendriks, P. & Hoeksema, J.) Dutch for non-native speakers; Level 2 (Language Centre)
Spring 2007-2008	Dutch for non-native speakers; Level 3 (Language Centre) Publishing using Word (Donald Smits Center for Information Technology) Presentation in English (Language Centre)
Autumn 2008-2009	Dutch for non-native speakers; Level 4 (Language Centre)



## Appendix B: Work plan

### Remainder of Year 2: November 2008-July 2009

#### *Research:*

Nov 2008	Writing article for the 6 <sup>th</sup> Anéla conference
Dec 2008	Corpus building: encyclopedia entries, fundraising letters
Dec 2008-Jan 2009	Writing paper for CiD post-conference proceedings
Dec 2008-Feb 2009	Corpus analysis (moves, coherence, cohesion): encyclopedia entries, fundraising letters
March 2009	Analysis of results
April 2009	Corpus building: popular scientific articles, advertisements
April-July 2009	Corpus analysis (moves, coherence, cohesion): popular scientific articles, advertisements

#### *Courses:*

17-20 Nov 2008	Excel (Donald Smits Center for Information Technology)
19-30 January 2009	LOT winter school, Groningen
Spring 2009	Methodology and statistics for linguistic research (Nerbonne, J.) Publishing in English (Language Centre) Dutch: Oral skills (Language Centre)
Jul-Aug 2009	?LSA Linguistic Institute (Berkeley, USA)

#### *Conference presentations:*

17-19 Dec 2008	Berzlánovich, I., Egg, M. & Redeker, G. Coherence and lexical cohesion in discourse from a genre perspective. <i>VIOT 2008: Taalbeheersing, the next level</i> , Amsterdam, 17- 19 December 2008.
27-29 May 2009	Berzlánovich, I., Egg, M. & Redeker, G. The interaction of coherence and lexical cohesion across genres. <i>6<sup>th</sup> Anéla Conference</i> , Kerkrade, The Netherlands.
June 2009	?TABU Day (Groningen, the Netherlands)

### Year 3: August 2009-July 2010

#### *Research:*

Sept 2009	Corpus analysis (move, coherence, cohesion): popular scientific articles, advertisements
Oct 2009	Analysis of results
Nov-Dec 2009	Preparing conference presentation(s) a/o writing a paper
Jan 2010	Corpus building: reviews
Jan 2010-April 2010	Corpus analysis (moves, coherence, cohesion): reviews
May 2010	Analysis of results
June-July 2010	Preparing conference presentation(s) a/o writing a paper

*Courses:*

Autumn 2009

Dutch: Writing skills (Language Centre)

Powerpoint advanced (Donald Smits Center for Information  
Technology)

*Conference presentations*

*Teaching:*

Spring 2010

**Year 4: August 2010-July 2011**

*Research:*

Dissertation

*Conference presentations*

*Teaching*