Maja Bärenfänger, Daniela Goecke, Mirco Hilbert, Harald Lüngen, Maik Stührenberg

Anaphora as an Indicator of Elaboration: A Corpus Study

This article describes an investigation of the relationship between anaphora and relational discourse structure, notably the ELABORATION relation known from theories like RST. A corpus was annotated on the levels of anaphoric structure and rhetorical structure. The statistical analysis of interrelations between the two annotation layers revealed correlations between specific subtypes of anaphora and ELABORATION, indicating that anaphora can function as a cue for ELABORATION.[1]

1 Introduction

Two aspects of the structure of discourse are *relational discourse structure* and *anaphoric structure*. There are two views regarding the relationship between these two levels of analysis: On the one hand, relational, hierarchical discourse structure is said to provide domains of accessibility for antecedent candidates of anaphoric expressions (Polanyi, 1988; Cristea et al., 2000; Asher and Lascarides, 2003). On the other hand, coreference plays a role in the definition of certain discourse relations, notably ELABORATION (Corston-Oliver, 1998; Carlson and Marcu, 2001; Knott et al., 2001), but also LIST e.g. in the discourse parsing approach by Corston-Oliver (1998, p. 137).

In an automated analysis of relational discourse structure of text, lexical discourse markers (i.e. conjunctions and sentence adverbials) play a major role as cues for identifying discourse relations (Marcu, 2000; Le Thanh et al., 2004). ELABORATION, however, is a discourse relation frequently not signalled by lexical discourse markers, hence the question arises whether one could systematically use anaphora as a cue for identifying ELABORATION. This study presents an empirical investigation of the relationship between discourse anaphora and relational discourse structure by means of an analysis of a text corpus that was annotated independently on these two levels of linguistic description. We focus on anaphoric structure as a cue for discourse structure, in particular, Elaboration. The remainder of the article is structured as follows: In Section 2, we provide the theoretical background of coreference and relational discourse structure as well as our categorial framework of anaphora and rhetorical relations and formulate our research questions in terms of these. In Section 3 we give an overview of our corpus of German scientific articles, the annotation schemes used for anaphora and rhetorical structure, and the methods used in querying and statistically analysing the corpus. In Section 4, the results of the corpus analysis are presented and discussed. In

[1] The work presented in this article is a joint effort of the projects A2 (*Sekimo*) and C1 (*SemDok*) of the Research Group 437 *Text-technological modelling of information* funded by the German Research Foundation DFG.

Section 5, we describe the implementation of some of our findings in a discourse parser, and present an evaluation of parsing experiments with and without anaphoric cues.[2]

2 Two aspects of discourse structure

2.1 Anaphora

Anaphoric relations as a cohesive device are an important factor of the coherence of texts. Anaphora occurs when the interpretation of a linguistic unit (the anaphor) is dependent on the interpretation of another element in the previous context (the antecedent). The anaphor is often an abbreviated or reformulated reference to its antecedent and thus provides for the progression of discourse topics. The analysis of anaphora as a device for discourse structure presupposes the notions of *discourse entities* and *discourse segments* (cf. Webber, 1988), the latter building the bridge to relational discourse structure.

Discourse entities – or discourse referents in the terminology of Karttunen (1976) – serve as constants within a discourse model which are evoked by (mainly) NPs and which can be referred to in the subsequent discourse. Following Webber (1988, p. 113), NPs can either evoke new discourse entities in the discourse model (or universe) or can "refer to ones that are already there". Pronouns do not evoke new discourse entities but access existing ones (cf. Webber, 1986). In DRT (Kamp and Reyle, 1993), a slightly different view on NPs evoking discourse referents is adopted. Each discourse is represented by a *discourse representation structure* (DRS), and each DRS consists of two components: a set of discourse referents (the *universe*) and a set of conditions. Both pronouns and NPs add discourse referents to the discourse universe and anaphoric relations to already existing referents are modelled via identity assertions whereas according to Webber (1986, 1988) an anaphoric relation holds directly by accessing already existing discourse entities.

For the investigation described in this article nominal discourse entities have been introduced for pronouns as well as for definite and indefinite NPs and anaphoric relations have been annotated manually on the basis of the discourse entities. Apart from anaphoric relations with antecedents of nominal type, anaphoric elements may also refer to antecedents that have been evoked by non-nominal units. Asher (1993, p. 35) uses the term *abstract entity anaphora* where "not just sentential nominals but other constructions like verb phrases or even whole sentences introduce abstract objects and eventualities into a discourse and may serve as referents for anaphoric pronouns".[3]

The following examples with nominal (1), sentential nominal (2), and verb phrase antecedents (3) illustrate the distinction between nominal and non-nominal discourse entities.

[2]We would like to thank the two anonymous reviewers who provided valuable comments on a previous version of this article.

[3]The term *sentential nominal* refers to constructions that are semantically related to sentential structures, e.g. due to a derived nominal as in Example (2) (cf. Asher, 1993).

(1) I met a man yesterday. He told me a story.
(Example taken from Clark, 1977, p. 414)

(2) [The destruction of the city]$_i$ amazed Fred. It$_i$ had been bloody.
(Example taken from Asher, 1993, p. 35)

(3) John saw [Mary cross the finish line first in the marathon]$_i$. Two days later, he still didn't believe it$_i$. (Example taken from Asher, 1993, p. 39)

The term discourse segment refers to either elementary spans of texts (clauses, sentences and the like) or complex segments that are built up recursively from elementary segments. Discourse segments and relations between them form the discourse structure which is of special interest for discourse anaphora; the interrelationship between anaphora and discourse structure is manifested in several approaches to discourse structure: Intentional approaches like Centering Theory (Grosz and Sidner, 1986; Grosz et al., 1995) model anaphora according to different relations between adjacent scentences. Informational approaches like SDRT (Segmented Discourse Representation Theory, Asher and Lascarides, 2003) model anaphoric relations on the basis of accessibility according to the underlying discourse structure. Discourse structure as a constraint for anaphoric relations is prominent in the *Right Frontier Constraint* (Polanyi, 1988). Furthermore, application-oriented approaches (e.g. Cristea et al., 1998, 2000) focus on the detection of appropriate antecedent candidates within an anaphora resolution system and use discourse structure as a constraint for anaphoric relations.

For a description of anaphoric relations one has to differentiate between the linguistic form of text spans between which anaphoric relations hold on the one hand and the semantic interpretations of the respective text spans, i.e. the discourse entities, on the other hand.

A taxonomy according to the linguistic form of the anaphoric element classifies anaphora into nominal anaphora, verb anaphora, adverb anaphora, zero anaphora and the like. Furthermore, the antecedent for nominal anaphors may be of nominal type or a non-nominal construction that refers to an abstract entity (e.g. events, facts, propositions; cf. Asher, 1993).

According to the relations that hold between the discourse entities, anaphora can be further divided into direct anaphora and indirect anaphora. For direct anaphora, the antecedent is explicitly mentioned in the previous context (Example (1) above) whereas for indirect anaphora the antecedent is not mentioned explicitly but has to be inferred from the context (Example (4)).

(4) I looked into the room. The ceiling was very high.
(Example taken from Clark, 1977, p. 415)

The latter is also referred to as *bridging relations* following the terminology of Clark (1977). Apart from the distinction of direct/indirect anaphora, discourse referents may be coreferent or not. In Example (1) the linguistic units "a man" and "he" are co-specified

and refer to the same entity whereas "the room" and "the ceiling" in Example (4) do not although they are closely related due to world knowledge.

The distinction of anaphora according (a) to the linguistic form of anaphor and antecedent and (b) to the relations that hold between anaphor and antecedent leads to a taxonomy of anaphoric relations consisting of two primary relations which can be used for a broad annotation and two sets of secondary relation types for a more fine-grained annotation. This taxonomy forms the basis for the annotation of anaphoric relations and has been defined, together with the annotation scheme, on the basis of Holler-Feldhaus (2004) and Holler et al. (2004). The annotation scheme is described in detail in Goecke et al. (2007) and in (Diewald et al., 2008, this volume). The primary relation types (cospecLink and bridgingLink) allow for a distinction of direct and indirect anaphora and may be further subdivided into secondary relation types according to the relation between anaphor and antecedent (see Figure 1).

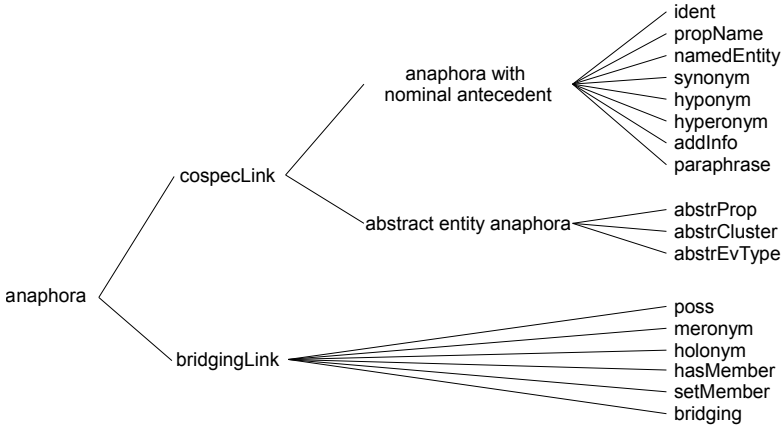


Figure 1: Sekimo hierarchy of anaphoric relations

For cospecLink two sets of secondary relations exist: one set for relations with antecedents of nominal type and one set for abstract entity anaphora. The subtypes of abstract entity anaphora are characterised as follows: abstrProp describes anaphoric relations with an antecedent of propositional type, abstEvType describes anaphoric relations with an event type antecedent, and abstrCluster describes anaphoric relations where the anaphor refers to a cluster of propositions. For nominal antecedents, we annotate eight secondary relation types: The relation ident is chosen for pronominal anaphors or anaphor-antecedent pairs with identical head noun. The value propName is chosen if the anaphoric element is a proper name that refers to an NP antecedent.

Anaphors that are not of type NAMEDENTITY but refer to an antecedent of type NAMEDENTITY are annotated with the respective relation type. Synonymy between the head nouns of anaphor and antecedent is annotated using the value SYNONYM. HYPERONYMY and HYPONYMY are chosen accordingly. The values ADDINFO and PARAPHRASE are chosen if the anaphor adds new information to the discourse or if the anaphor is a paraphrase of its antecedent.

For bridging relations six secondary relation types have been defined: The value POSS describes a possession relation between the anaphor and its antecedent. The value MERONYM is chosen in case of a part-whole-relation between the head nouns of anaphora and antecedent; HOLONYM is chosen accordingly. The value HASMEMBER is chosen if the anaphor describes a set and the antecedent(s) are part of of that set and SETMEMBER is chosen if the anaphoric elements is part of a set described by its antecedent. If none of the previous relation types hold the relation is annotated using the value BRIDGING.

The taxonomy shows that not only pronominal anaphors or definite descriptions with identical head nouns are taken into account for the investigation of anaphora and relational discourse structure. The majority of the relation types are relevant for definite description anaphors whose relations are licensed by lexical-semantic relations or association (e.g. *birthday party - presents*). Both intra- as well as inter-sentential anaphora is taken into account; definite description anaphors tend to find their antecedents across sentence boundaries even at a large distance between anaphor and antecedent. Consequently, anaphor and antecedet are frequently located within different discourse segments, allowing for an an investigation of the relationship between discourse anaphora and relational discourse structure.

The applicability of the taxonomy for corpus annotations has been tested in a study on inter-annotator agreement. The results of the study show that annotators are able to annotate even fine-grained secondary relation types reliably (cf. Goecke et al., 2008).

2.2 Rhetorical structure

Relational discourse structure is covered by several linguistic theories of discourse like SDRT, the Unified Linguistic Discourse Model (ULDM, Polanyi, 1988; Polanyi et al., 2003), or Rhetorical Structure Theory (RST, Mann and Thompson, 1988; Marcu, 2000). In the framework of RST, which we focus on here, discourse structure consists of relationally connected discourse segments which can be either elementary or complex. Segments are combined to form larger segments by two types of discourse relations: mononuclear or multinuclear relations. In a mononuclear relation, one discourse segment has the status of a "nucleus" (N), the more "essential" piece of text, the other segment has the status of a "satellite" (S), a less essential text part "more suitable for substitution" (cf. Mann and Thompson, 1988). In a multinuclear relation, all related segments serve as nuclei. The original RST distinguishes 26 mono- or multinuclear relations; like other projects (cf. Carlson et al., 2001; Hovy and Maier, 1995), we extended this relation set

with subrelations according to requirements of our corpus and application scenario (cf. Lüngen et al., 2006).

One prominent relation in our corpus is the mononuclear relation ELABORATION. Mann and Thompson (1988) introduced ELABORATION into RST by defining conditions on the combination of two discourse segments S and N for ELABORATION to hold:

> S presents additional detail about the situation or some element of subject matter which is presented in N or inferentially accessible in N in one or more of the ways listed below. In the list, if N presents the first member of any pair, then S includes the second:
> 1. set:member
> 2. abstract:instance
> 3. whole:part
> 4. process:step
> 5. object:attribute
> 6. generalization:specific
>
> (*ibid.* p. 273).

The relations enumerated in this listing partly resemble the semantic relations introduced in Section 2.1. The use of "presents", "presented in" and "includes" in the definition suggests that the relations listed are supposed to hold between entities that are in a sense contained in the segments N and S.

Corston-Oliver (1998, p. 81), who focuses on discourse parsing, argues that ELABORATION is amongst other things indicated by "subject continuity" which he describes as being "the most important kind of referential continuity for identifying discourse relations". In his "worked example" (*ibid.* p. 203f), cf. Example (5), subject continuity is clearly realised by the anaphoric pronoun *it*, and subject continuity also appears in his list of cues for ELABORATION (*ibid.* p. 103).

(5) [The aardwolf is classified as Proteles cristatus]$_{Nuc}$. [It is usually placed in the hyena family, Hyaenidae. {...}]$_{Sat}$

(Example taken from Corston-Oliver, 1998, p. 203f; originally from an article in the *Microsoft Encarta 96* Encyclopedia)

Wolf and Gibson (2006, p. 32) also use an ELABORATION relation in their discourse annotation schema (which is not based on RST) and define it in their coding procedure as providing "more detail about an already introduced entity or event".

(6) [Crawford & Co., Atlanta (CFD) began trading today]$_{Nuc}$. [Crawford evaluates health care plans, manages medical and disability aspects of worker's compensation injuries and is involved in claims adjustments for insurance companies.]$_{Sat}$

(Coding example in Wolf and Gibson, 2006, p. 32f; originally from text wsj-0607 (Wall Street Journal Corpus) from Harman and Liberman (1993))

In their coding example (Example (6)), the discourse entity named *Crawford* is referred to by linguistic expressions in both segments. Wolf and Gibson (2006) do not claim that

anaphora is a (necessary) criterion for ELABORATION. They formulate more generally that "[o]ften when there is an anaphoric relation between two discourse segments, these discourse segments are also related by a coherence relation" (p. 35).

ELABORATION relations have also been compared to focus structures (Knott et al., 2001) such as described in Centering Theory (Grosz and Sidner, 1986; Grosz et al., 1995), which models anaphora across adjacent sentences.

Though in none of the definitions cited above it is explicitly said that in an ELABORATION relation between two discourse segments, a discourse entity or referent in N is continued in S by a co-specified linguistic expression, it is the case in many examples that we found in the literature including those presented above. Terms like "situation", "element of subject matter", "subject", "entity", and "event" seem to refer to different types of discourse entities.

Because of this frequent association of ELABORATION with semantic relations between certain distinguished discourse entities, we also believe that it can be compared to types of "thematic progression" or "thematic development" known from text linguistics. The following is a simplified description of types of thematic progression as introduced in Daneš (1970) and Zifonun et al. (1997):

1. Continuation of theme or rheme
2. Derivation or integration from the preceding theme or hypertheme
    a) derivation from hypertheme
    b) derivation from preceding theme or rheme
    c) integration of preceding themes in one hypertheme

Thematic relations between segments with a common topic abound in any given text, and according to (Carlson et al., 2001, p. 53) ELABORATION is "extremely common at all levels of the discourse structure" as well. In our corpus, it is the most frequent relation (43% of all relations in the SemDok-corpus and 38% of all relations in the subcorpus used for the analyses described in this article, see Section 3.1). ELABORATION is much less constrained than most other RST relations and seems to be a natural "default relation" to be assigned when no other relation can be assigned due to an absence of lexical discourse markers (another candidate for a default relation is LIST).

In order to render the original RST-definition of ELABORATION by (Mann and Thompson, 1988, p. 273) more detailed, we extended the set of rhetorical relations for our annotation project with subtypes of ELABORATION and with definitions which make reference to discourse entities and themes. In doing so, we also compared other sets of subtypes of ELABORATION found in the research literature, i.e. Mann and Thompson (1988), Hovy and Maier (1995), and Carlson et al. (2001), with relation instances in our corpus labelled as ELABORATION. The hierarchy of ELABORATION relations used in the final version of our corpus annotation scheme is shown in Figure 2. In the annotation of the sample corpus described in Section 3.3 , annotators where asked to use only the terminal types, i.e. the leaves of the hierarchy for annotation, except for those types that are marked with an asterisk '*' in Figure 2. Only if annotators definitely could

not decide on one terminal type were they allowed to annotate one of the intermediate types.

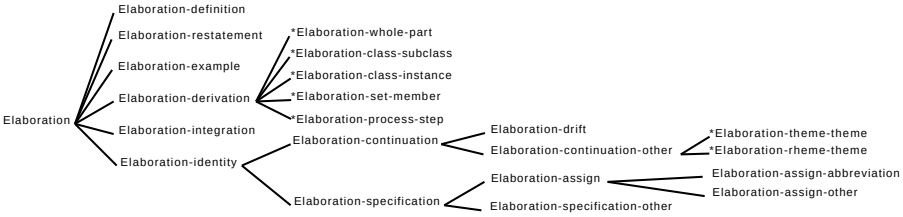


Figure 2: SemDok hierarchy of ELABORATION relations

The first three subrelations ELABORATION-DEFINITION, ELABORATION-RESTATEMENT, and ELABORATION-EXAMPLE are not defined in terms of thematic progession or referential continuation, but rather along the lines of the relations DEFINITION, EXAMPLE, and RESTATEMENT in Carlson et al. (2001), and in the annotation task, they take priority over an assignment of one of the remaining subtypes. ELABORATION-DEFINITION holds when the satellite contains a definition of a technical concept occurring in the nucleus. In our corpus, it is frequently signalled by a colon terminating the nucleus, and/or XML markup such as the DocBook <`glossentry`> element on the annotation layer of logical document structure (cf. Walsh and Muellner, 1999). ELABORATION-EXAMPLE holds, when the satellite represents an example of the nucleus or of a concept in the nucleus. It is generally accompanied by a lexical discourse marker in the satellite such as *z.B.* or *beispielsweise* (cf. Example (7)). Finally, ELABORATION-RESTATEMENT holds, when the satellite represents a reformulation of the nucleus of about the same length.

The subrelation ELABORATION-IDENTITY, on the other hand, is characterised by a thematic or referential identity between nucleus and satellite. In case of its subtype ELABORATION-CONTINUATION, there is thematic continuity between nucleus and satellite either in the form of a common hypertheme (subtype ELABORATION-DRIFT) or in the form of an explicit linguistic expression in the satellite that refers to the rheme or theme of the nucleus (subtype ELABORATION-CONTINUATION-OTHER[4]). ELABORATION-DRIFT is further defined to cover the following cases: a.) The hypertheme need not necessarily be mentioned in the nucleus or satellite, but it should be nameable, b.) a theme that was introduced in the nucleus as an NP is continued in the satellite in an embedded phrase only (cf. Example (11)), or c.) a thematic event anaphor (like *dies*) in the satellite refers to the proposition or set of propositions that forms the nucleus (cf. Example (10)).

In case of the other subtype of ELABORATION-IDENTITY, ELABORATION-SPECIFICATION, the satellite is about the same discourse entity in such a way that the meaning of

[4]The suffix "-other" was used to distinguish the major subtype of ELABORATION-CONTINUATION, ELABORATION-SPECIFICATION, and ELABORATION-ASSIGN from its co-subtypes, respectively.

the nucleus is extended, restricted or further specified by a modifying phrase only, i.e. as an incomplete sentence and without explicitly mentioning the thematic discourse entity again. Its subtype ELABORATION-ASSIGN holds, when the meaning of the nucleus is in a way *assigned* by the author to the expression in the satellite. In academic texts, this frequently occurs when abbreviations or acronyms are introduced (subtype ELABORATION-ASSIGN-ABBREV). ELABORATION-ASSIGN is thus similar to ELABORATION-DEFINITION, but with inverse nuclearity. The regular instances of ELABORATION-SPECIFICATION which are not covered by ELABORATION-ASSIGN, are labelled ELABORATION-SPECIFICATON-OTHER (see Example (9)).

(7) ELABORATION-EXAMPLE:[5] [*Åland hat auch in vielen anderen Hinsichten eigene Gesetze,*]$_{Nuc}$ [*z.B. sind die Inseln entmilitarisiert.*]$_{Sat}$

(8) ELABORATION-CONTINUATION-OTHER:[6] [*Im folgenden Abschnitt werden wir zunächst einige terminologische Klärungen vornehmen.*]$_{Nuc}$ [*Diese betreffen einerseits unser Verständnis von regionalen Varietäten (2.1), andererseits das Spracheinstellungskonzept (2.2).*]$_{Sat}$

(9) ELABORATION-SPECIFICATION-OTHER:[7] [*Ob regionale Varietäten* [*(Dialekte, Regionalsprachen, nationale Standardvarietäten)*]$_{Sat}$ *Thema des Deutsch als Fremdsprache-Unterrichts sein können bzw. sein sollten, ist in den letzten Jahren zunehmend zum Gegenstand kontroverser Diskussionen geworden.*]$_{Nuc}$

(10) ELABORATION-DRIFT:[8] [*Die vorherrschende Meinung insbesondere bei DaF-Lehrern und bei den meisten Lehrbuchverlagen scheint zu sein, dass sich der DaF-Unterricht hauptsächlich auf die Vermittlung der deutschen Standardsprache beschränken muss und soll.*]$_{Nuc}$. [*Dies spiegelt sich zum einen in der Vernachlässigung regionaler Varietäten in DaF-Lehrwerken zugunsten der Standardsprache wider {...}*]$_{Sat}$

(11) ELABORATION-DRIFT:[9] [*Automatisierte Prozesse im L2-Erwerb sind solche, auf die keine oder nur geringe Aufmerksamkeit gerichtet wird.*]$_{Nuc}$ [*Eine wichtige Funktion der Automatisierung ist die Freisetzung von Kapazitäten für die gleichzeitige Bewältigung von aufmerksamkeitsintensiven Aktivitäten.*]$_{Sat}$

(12) ELABORATION-DERIVATION:[10] [*Die Erhebung und Analyse der mündlichen Primärdaten erfolgt in zwei großen Blöcken.*]$_{Nuc}$ [*In einer Querschnittsuntersuchung wird zunächst die Frage untersucht, wie {...}. Hiervon ausgehend können im zweiten Block longitudinal Veränderung von {...} verfolgt werden.*]$_{Sat}$

ELABORATION-DERIVATION, which is another direct subtype of ELABORATION, is

[5] Example taken from Mirja Saari (2000): "Schwedisch als die zweite Nationalsprache Finnlands: Soziolinguistische Aspekte". In: *Linguistik Online 7*, http://www.linguistik-online.de.

[6] Example taken from Harald Baßler, Helmut Spiekermann (2001): "Dialekt und Standardsprache im DaF-Unterricht. Wie Schüler urteilen - wie Lehrer urteilen". In: *Linguistik Online 9*, http://www.linguistik-online.de.

[7] Example taken from Baßler/Spiekermann (2001).

[8] Example taken from Baßler/Spiekermann (2001).

[9] Example taken from Olaf Bärenfänger, Sabine Beyer (2001): "Zur Funktion der mündlichen L2-Produktion und zu den damit verbundene kognitiven Prozessesn für den Erwerb der fremdsprachlichen Sprechfertigkeit". In: *Linguistik Online 8*, http://www.linguistik-online.de.

[10] Example taken from Bärenfänger/Beyer (2001).

based on thematic derivation, i.e. comprises whole-part, class-subclass, class-instance, set-member, or process-step relations between entities in the nucleus and the satellite (cf. Example (12)). ELABORATION-INTEGRATION is its opposite, with the inverse relation pairs, i.e. part-whole, subclass-class etc.

Only few of our subrelations are accompanied by explicit lexical, grammatical, or punctuational discourse markers, e.g. ELABORATION-EXAMPLE (*z.B.*) or ELABORATION-SPECIFICATION (parenthesis and phrase status of satellites), but the most frequently occurring subtypes of ELABORATION are not signalled by explicit discourse markers and cannot automatically be determined on the basis of lexical or grammatical cues.

2.3 Research questions and hypotheses

Based on our understanding of ELABORATION as indicating thematic relations in the framework of RST, it seems reasonable to look for the cues that are also used for the analysis of thematic relations. One prominent signal of thematic connections are referential ties between adjacent sentences, or more specifically: references between sentence themes (cf. Daneš, 1970; Givón, 1983, 1992). Sentence themes are signalled by nominal discourse entities, often expressed as pronouns, definite NPs, NPs in sentence initial position, or NPs in the role of grammatical subject. Anaphoric relations between adjacent discourse segments should therefore be good indicators for thematic relations, and hence for ELABORATION. Figures 3 and 4 exemplify this interrelationship: In the former figure, the cue for the discourse relation is a lexical discourse marker whereas in the latter figure, the discourse relation has an anaphoric relation as its cue.

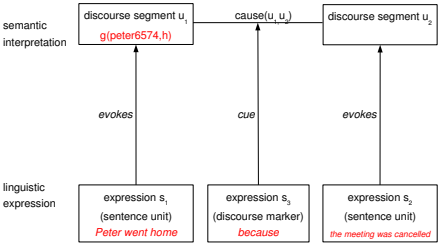
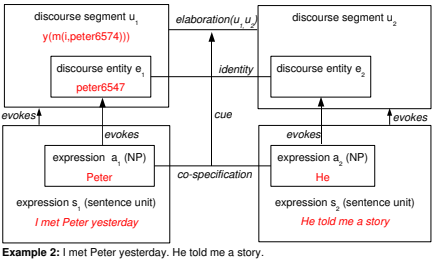


Figure 3: Linguistic expressions and semantic interpretation with lexical discourse marker

Generally, we expect that anaphora is a *necessary* condition for Elaboration while we also want to test whether it could be a *sufficient* condition. Furthermore, we expect that specific anaphoric relations from the scheme introduced in Section 2.1 correspond to specific ELABORATION relations; we would, for example, expect that ELABORATION-CONTINUATION-OTHER is indicated by the anaphoric relation *cospec:ident*. An overview of expected correspondences between thematic relations, ELABORATION relations and



Example 2: I met Peter yesterday. He told me a story.

Figure 4: Linguistic expressions and semantic interpretation with anaphora

anaphoric relations is given in Table 1. (Only those ELABORATION subrelations for which we had expectations are shown.).

Table 1: Theoretical correspondences between thematic relations, ELABORATION relations and anaphoric relations

Thematic Relations	**Elaboration Relations**	**Anaphoric Relations**
Continuation of theme or rheme	ELABORATION-CONTINUATION-OTHER	*cospec:synonym* *cospec:paraphrase* *cospec:ident*
Derivation from preceding theme or rheme	ELABORATION-DERIVATION	*cospec:hyperonym* *bridging:holonym* *bridging:setMember*
Integration of preceding theme or themes in one hypertheme	ELABORATION-INTEGRATION	*bridging:meronym* *cospec:hyponym* *bridging:hasMember*
Derivation from hypertheme	ELABORATION-DRIFT	*bridging:bridging* *bridging:poss* *bridging:abstrProp* *bridging:abstrCluster*

Since ELABORATION-CONTINUATION-OTHER should have been annotated when an explicit linguistic expression refers to the theme or rheme of the nucleus, we would expect it to be accompanied by semantic relations between discourse entities that indicate referential identity, i.e. *cospec:synonym*, *cospec:paraphrase*, or *cospec:ident*. Since ELABORATION-DERIVATION is based on *whole-part*, *class-subclass*, *class-instance* *set-member*, *process-step* (terms from the definition by Mann and Thompson, 1988) relations between entities in nucleus and satellite, we would expect it to be accompanied by the semantic relations *cospec:hyperonym*, *bridging:holonym*, or *bridging:setMember*. Since ELABORATION-INTEGRATION is the opposite we would expect it to appear together with *bridging:meronym*, *cospec:hyponym*, or*bridging:hasMember*. As *Elaboration-drift* may hold due to a common hypertheme, it may firstly appear together with *bridg-*

ing:bridging; since it may hold on account of thematic continuation realised in an embedded phrase, it may secondly be accompanied by *bridging:poss* (In an NP like *seine Untersuchung*, the possessive pronoun takes the position of an NP in genitive which is embedded in the whole, higher-level NP as a whole whose head is *Untersuchung*). Thirdly, since ELABORATION-DRIFT may hold when a thematic continuation is realised by an event anaphor, it may be accompanied by a *bridging:abstrProp* or *bridging:abstrCluster* relation.

Since non-thematic anaphoric relations between discourse segments might theoretically hold as well, one research question is whether the theoretical correspondences in Table 1 work as practical indicators of ELABORATION. Our general goal is to investigate inhowfar our theoretically derived claims are supported by empirical evidence by analysing a corpus that has been annoted on the level of anaphoric structure and on the level of rhetorical structure.

Our corpus, its relevant linguistic annotations and the analysis tools are described in the following section.

3 Methods

3.1 Corpus

The SemDok corpus used both for research on discourse structure and anaphoric structure consists of 47 German linguistic scientific journal articles, formally annotated on the levels of syntax, morphology and document structure. For the analysis of correlations between anaphoric relations and ELABORATION relations we developed a sample corpus, which comprises two scientific journal articles from the SemDok corpus, one web-published scientific article and one newspaper article (altogether 15.622 word forms). These four texts were segmented in elementary and complex discourse segments, and annotated on the levels of rhetorical structure (RST-HP, for RST, hypotaxis and parataxis) and discourse entities and anaphoric relations (CHS, for cohesion). The two kinds of annotations have been carried out independently.

3.2 Annotation of anaphoric structure

The corpus under investigation has been annotated manually for anaphoric relations using the annotation tool *Serengeti* which is described in detail together with the annotation scheme in Diewald et al. (2008, this volume). Anaphoric relations are marked between text spans, i.e. between linguistic units (*markables*). These text spans evoke discourse entities as part of the discourse universe, thus anaphoric relations are *marked* between linguistic units but the corresponding semantic relations *hold* between discourse entities.

Each text of the corpus has been preprocessed using the dependency parser Machinese Syntax[11] which provides lemmatisation, POS information, dependency structure, mor-

[11] `http://www.connexor.eu`.

phological information and grammatical function. Based on this information, markables of nominal type have been detected automatically by identifying nominal heads (i.e. nouns or pronouns) and their premodifiers.

The annotation procedure has been performed in two steps. The first step has been done before the data analysis and its focus lay on the annotation of anaphoric relations between nominal anaphors and antecedents of nominal type only. The second annotation step has been done after the analysis of the nominal data of step 1 and its focus lay on the annotation of abstract entity anaphora including adverb anaphors (cf. Figure 1).

During the first step, a complete annotation has been done for those anaphoric relations with both anaphor and antecedent of nominal type. These relations include pronominal anaphors as well as definite description anaphors with nominal antecedents where both intra- as well as inter-sentential anaphora has been taken into account.

In a second step, abstract entity anaphora has been annotated. These relations hold between the anaphor and an antecedent of propositional or event type. Whereas the first step has been a complete annotation of both markables and anaphoric relations the second step has been a partial annotation, only. Due to the vast amount of all propositions and events in a text, only those discourse entities have been identified as markables that form the antecedent of an anaphoric relation. Three types of discourse entities have been annotated manually: The type CLUSTER describes discourse entities that are evoked by several adjacent sentences, PROP describes entities evoked by one proposition (sentence or embedded clause), and EVTYPE describes all entities evoked by a verb and its arguments. Furthermore, all adverbial anaphors (such as *hierbei*, *dabei*) have been marked as discourse entity of type ADV in order to annotate adverb anaphora leading to a total number of five different types of discourse entities: NOMINAL, ADV, PROP, EVTYPE, and CLUSTER.

For the corpus under investigation a total number of 662 anaphoric relations has been annotated during the first step; during the second step another 68 abstract entity anaphors have been annotated.

3.3 Annotation of rhetorical structure

Rhetorical structure according to RST was encoded in the XML application RST-HP developed in the project SemDok (Lüngen et al., 2008). Discourse segments are marked using the two elements `hypo` and `para` with a relation name in the `@relname` attribute (see Lüngen et al., 2006, 2008, for a description and sample annotations of RST-HP). Unlike URML (Reitter and Stede, 2003) and the XML-like format put out by the RSTTool (O'Donnell, 2000), RST-HP exploits the XML document tree to represent an RST tree, which means that general XML query tools such as XPath or the Sekimo Tools (Witt et al., 2005) can be applied straightforwardly to query RST-HP annotations.

In manual or automatic annotation, rhetorical relations are assigned on the basis of the *RRSet*, a taxonomy comprising 70 rhetorical relation types for the analysis of the discourse structure of scientific articles, 44 of which are base types to be used in the manual and automatic annotations (Bärenfänger et al., 2008). The ELABORATION

sub-hierarchy given in Figure 2 is part of the RRSet taxonymy. The annotation guidelines stated that when a lexical discourse marker for an ordinary relation could be found, this relation should be annotated while the conditions for ELABORATION need not be checked. This procedure, which as we think is typical for RST analyses, gives ELABORATION the status of a default relation.

The RST annotation of the four articles of the sample corpus was done using O'Donnell's RSTTool. The XML-like format that is output by the RSTTool was converted to RST-HP by means of a perl program. Each file was annotated independently by two annotators, who then discussed possible annotation differences and agreed on a single "master" version which was subsequently used in the comparison with annotations of anaphoric structure described below.

For the present study, we concentrated on subtrees of RST trees for complex discourse segments of type "block", i.e. trees where the minimal units are *elementary discourse segments* (basically clause-like units) and whose root node corresponds to a paragraph. The RST-HP annotations for block segments constructed in the sample corpus contained 846 RST subtrees altogether.

To get an idea of inter-annotator agreement for the RST relation assignment task, we measured agreement within "block" segments for three articles that were coded by three annotators each. Kappa values for the nine resulting annotator pairings ranged between 0.47 and 0.81 which is interpreted as 'moderate agreement' to 'almost perfect agreement' by Landis and Koch (1977).

3.4 Analysis

During the annotation of anaphoric and rhetorical structure, the primary data of the input documents were left unchanged so that the Sekimo query tools could be employed for querying relations between elements of two XML annotation layers (cf. Witt et al., 2005). We focused on the analysis of the *inclusion* relation to verify whether a discourse entity on the CHS layer was included in a discourse segment on the RST-HP annotation layer.

In order to research the hypotheses formulated in Section 2.3, we firstly derived the set of instances of adjacent discourse segments DS_i and DS_j that contained an anaphoric expression in DS_j whose antecendent was contained in DS_i, together with the information of whether DS_i and DS_j formed a combined RST subtree in RST-HP with a relation assignment or not. This query resulted in an XML dataset of 662 anaphoric instances. Secondly, we derived the set of instances of adjacent discourse segments that formed a combined RST subtree in RST-HP, together with the information about an occurrence of anaphora formed by an anaphoric expression in DS_j and a related antecedent in DS_i, and if applicable, its type. This query resulted in an XML dataset of 846 relation instances. To obtain the statistics reported in Section 4, these two databases were queried using XPath expressions.

4 Results and Discussion

4.1 Is anaphora a sufficient condition for ELABORATION?

That the existence of an anaphoric relation might not be a sufficient condition for the discourse relation of ELABORATION to hold seems obvious as anaphora can also be involved in other relations. Most other relations are defined without recourse to referential structure or thematic progression, and are frequently signalled by a lexial discourse marker. But in order to quantify the degree in which anaphora might or might not be a sufficient condition for ELABORATION, we checked all anaphoric instances for their co-occurrence with ELABORATION. The results of this investigation are given in Table 2.

Table 2: Is anaphora a sufficient condition?

	Total No.
anaphoricInstance	662
@rtype='elaboration'	176
@rtype='no-RST-relation'	301
@rtype='RST-relation-other-than-elaboration'	185

Due to the fact that ELABORATION is a default relation, we had expected anaphoric relations to coincide with relations other than ELABORATION: 185 out of 662 anaphoric instances (27,95%) coincide with relations other than ELABORATION whereas 176 anaphoric instances (26.59%) coincide with ELABORATION.

Interestingly, the majority of anaphoric instances (45,47%) does not coincide with any relation at all. These instances are either located within the same discourse segment or there is no rhetorical relation between the relevant segments due to the overall discourse segmentation.

Clearly, the occurrence of an anaphoric relation is not a sufficient condition for ELABORATION. In the following section we will investigate the question whether the existence of an anaphoric relation is a *necessary condition* for ELABORATION.

4.2 Is anaphora a necessary condition for ELABORATION?

In the corpus, 298 ELABORATION instances could be identified on the basis of the first annotation step, but for only 191 of them, an anaphoric relation holds between discourse entities in the related discourse segments. In 107 cases, ELABORATION does not correlate with an anaphoric relation (Table 3). The different subtypes of ELABORATION deviate with respect to the strength of their interrelation with anaphoric relations. ELABORATION-CONTINUATION-OTHER correlates almost always with an anaphoric relation, whereas ELABORATION-SPECIFICATION-OTHER and ELABORATION-ASSIGN-OTHER are only weakly associated with anaphoric relations – for them, there are more occurrences without an anaphoric relation present than with an anaphoric relation. How can these differences be explained?

Firstly, there is the technical reason that in the definitions of ELABORATION-SPECIFICATION-OTHER and ELABORATION-ASSIGN-OTHER, the satellite is described to have phrasal status (i.e. not clausal), and such units mostly correspond to parenthetical segments. Anaphoric relations to discourse entitities in parentheses, however, were not marked on the CHS annotation layer.

Secondly, neither ELABORATON-SPECIFICATION-OTHER, ELABORATION-ASSIGN-OTHER, ELABORATION-DEFINITION, nor ELABORATION-EXAMPLE are typical thematic continuations or derivations. Instead of being signalled by referential ties, they are indicated by lexical and syntactic cues (cf. Section 2.2): ELABORATION-EXAMPLE is almost always marked by lexical markers like "z.B." or "beispielsweise", and ELABORATION-SPECIFICATION-OTHER and ELABORATION-ASSIGN-OTHER are indicated by parentheses or brackets which encloses the NPs or PPs in the satellite that specifies, extends or restricts an entity in the nucleus without repeating the entity itself.

Another relation which shows a different behaviour than expected is ELABORATION-DRIFT. Although this relation is defined as exhibiting some sort of thematic continuity, it does not – like ELABORATION-CONTINUATION-OTHER – frequently correlate with anaphoric relations (see Table 3). 44 out of 111 instances of ELABORATION-DRIFT are not connected by an anaphoric relation at all. These result was so much against our expectations that we decided to carry out a qualitative analysis of the 107 ELABORATION instances which had no correspondence with an anaphoric relation.

Table 3: Number of ELABORATION instances with anaphoric relations

	All	**With anaphoric relations**		**Without anaphoric relations**	
ELABORATION-DRIFT	111	67	(60.36%)	44	(39.64%)
ELABORATION-CONTINUATION-OTHER	56	51	(91.07%)	5	(8.93%)
ELABORATION-SPECIFICATION-OTHER	43	20	(46,51%)	23	(53,49%)
ELABORATION-DERIVATION	36	28	(77.78%)	8	(22.22%)
ELABORATION-DEFINITION	13	6	(46,15%)	7	(53,85%)
ELABORATION-EXAMPLE	10	4	(40%)	6	(60%)
ELABORATION-INTEGRATION	7	4	(57,14%)	3	(42,86%)
ELABORATION-IDENTITY	7	4	(57,14%)	3	(42,86%)
ELABORATION-ASSIGN-OTHER	6	1	(16,67%)	5	(83,33%)
ELABORATION	5	4	(80.00%)	1	(20.00%)
ELABORATION-RESTATEMENT	3	1	(33,33%)	2	(66,67%)
ELABORATION-CONTINUATION	1	1	(100.00%)	0	(0.00%)
All Elaboration-Trees	298	191	(64,09%)	107	(35,91%)

The qualitative analysis showed that a bulk of the missing anaphoric relations were due to the scope of the anaphoric relation set and the annotation focus chosen in the project Sekimo, which was on nominal antecedents only. Propositional antecedents had not been taken into account during the first annotation phase. In 37 of the 107 not anaphorically linked ELABORATION instances, anaphoric relations could – according to the findings of the qualitative analysis – be established on the basis of a propositional

antecedent. These abstract entity anaphors were then annotated in a second annotation step.

For another 38 instances it was possible to assign types of anaphoric relations that are not based on lexical-semantic relations, but involved other, e.g. morpho-semantic relations (e.g. derivation) or broad association (such as *Kind – Infantilisierung* in the sample corpus). Whereas anaphora due to identity of head nouns or due to lexical-semantic relations can be decided rather unambiguously, this is not the case for anaphora based on association. Narrow association (e.g. *wedding – bride*) is detected more easily than broad association. But taking broad association into account helped to identify additional anaphoric relation instances such that subsequently only six instances (i.e. 2,69%) of the 223 instances related by ELABORATION-CONTINUATION-OTHER, ELABORATION-DRIFT, ELABORATION-DERIVATION, ELABORATION-INTEGRATION, and ELABORATION-DEFINITION had no anaphoric connection. Table 4 shows the effect of the qualitative analysis as well as of the second annotation step.

Table 4: ELABORATION instances with anaphoric relations after qualitative analysis and second annotation step

	All	**With anaphoric relations**		**Without anaphoric relations**	
ELABORATION-DRIFT	111	108	(97.30%)	3	(2.70%)
ELABORATION-CONTINUATION-OTHER	56	55	(98.21%)	1	(1.79%)
ELABORATION-SPECIFICATION-OTHER	43	25	(58.14%)	18	(41.86%)
ELABORATION-DERIVATION	36	35	(97.22%)	1	(2.78%)
ELABORATION-DEFINITION	13	12	(92.31%)	1	(7.69%)
ELABORATION-EXAMPLE	10	7	(70.00%)	3	(30.00%)
ELABORATION-INTEGRATION	7	7	(100%)	0	(0%)
ELABORATION-IDENTITY	7	7	(100%)	0	(0%)
ELABORATION-ASSIGN-OTHER	6	1	(16.67%)	5	(83.33%)
ELABORATION	5	5	(100%)	0	(0%)
ELABORATION-RESTATEMENT	3	3	(100%)	0	(0%)
ELABORATION-CONTINUATION	1	1	(100%)	0	(0%)
All Elaboration-Trees	298	266	(89.26%)	32	(10.74%)

Altogether, the revised quantitative analysis of the correlations between ELABORATION and anaphoric relations shows that 108 of 111 instances of ELABORATION-DRIFT, 35 out of 36 instances of ELABORATION-DERIVATION, seven out of seven instances of ELABORATION-INTEGRATION and 55 out of 56 instances of ELABORATION-CONTINUATION-OTHER indeed co-occur with an anaphoric relation. Only the figures for ELABORATION-SPECIFICATION-OTHER, ELABORATION-ASSIGN-OTHER and ELABORATION-EXAMPLE did not differ significantly after the qualitative analysis. Our second hypothesis – that an anaphoric relation is a necessary condition for ELABORATION – must therefore be considered true for all subtypes of ELABORATION except ELABORATION-SPECIFICATION-OTHER, ELABORATION-ASSIGN-OTHER and ELABORATION-EXAMPLE. Note that the latter three relations comprise the majority of cases where ELABORATION is marked by a lexical discourse marker or by parenthesis.

In Table 1 we pointed out that specific subtypes of ELABORATION are expected to correspond to specific thematic relations and anaphoric relations. The hypothesised correspondences could be partly supported by the quantitative analysis of the corpus. The results differed with respect to their relative frequency. Stronger correlations with certain anaphora types were found for ELABORATION-CONTINUATION-OTHER, ELABORATION-DERIVATION and ELABORATION-INTEGRATION. The most frequent anaphoric relations contained after the first annotation step are shown in Table 5.[12]

Table 5: Co-occurrences of ELABORATION relations and anaphoric relations

Elaboration Instances	**With Anaphoric Relations Contained**
ELABORATION-CONTINUATION-OTHER 56 (51 with anaphora)	38 (63) x *cospec:ident* 6 (15) x *bridging:setMember* 6 (8) x *cospec:paraphrase* 6 (7) x *bridging:bridging* 3 (5) x *bridging:poss* 3 (3) x *cospec:synonym*
ELABORATION-DERIVATION 36 (28 with anaphora)	17 (43) x *bridging:setMember* 13 (17) x *cospec:ident* 3 (5) x *bridging:bridging* 3 (5) x *cospec:isA* 2 (4) x *cospec:synonym*
ELABORATION-INTEGRATION 7 (4 with anaphora)	2 (3) x *bridging:hasMember* 2 (2) x *cospec:paraphrase*
ELABORATION-DRIFT 111 (67 with anaphora)	41 (60) x *cospec:ident* 12 (13) x *bridging:bridging* 11 (13) x *cospec:paraphrase* 10 (17) x *bridging:setMember*

ELABORATION-CONTINUATION-OTHER co-occurs with *cospec:ident* most of the time (38 of 56 cases, i.e. 67.86% of all instances of ELABORATION-CONTINUATION-OTHER co-occur with *cospec:ident*), six co-occur with *cospec:paraphrase*. 17 of 36 instances of ELABORATION-DERIVATION co-occurred with *bridging:setMember*, and two of three instances of ELABORATION-INTEGRATION co-occurred with *bridging:hasMember*. By contrast, the findings for ELABORATION-DRIFT were much more ambiguous: It co-occurs with *cospec:ident* (41 of 111 instances), *cospec:paraphrase* (eleven of 111 instances), *bridging:bridging* (twelve of 111 instances) and *bridging:setMember* (ten of 111 instances). Despite these ambiguities, some types of anaphoric relations might help automatically identify a specific ELABORATION relation when no other rhetorical relations can be determined, and we report a test of this in Section 5.

The qualitative analysis of the corpus also suggested that anaphoric expressions that correlated with ELABORATION are more frequently found in sentence-initial position

[12] In the column entitled 'with anaphoric relations contained', the first figure represents the number of ELABORATION instances that contain at least one anaphoric instance of the type, and the second figure in brackets represents the total number of anaphoric instances of the type contained

(*vorfeld*) or in the role of the grammatical subject (e.g. in 42 of the 51 ELABORATION-CONTINUATION-OTHER instances with anaphora) than in a different position or role. This is presumably due to the fact that subject and vorfeld positions are typical *topic* (i.e. sentence theme) positions in German syntax.

5 Discourse parsing experiments

In order to evaluate the contribution of an analysis of anaphora to automated discourse parsing, we integrated a processing of anaphoric cues from the CHS annotation layer of an input document in the RST-based discourse parser developed in the SemDok project.

The central component of the parsing system is called GAP – Generalised Annotation Parser. GAP is a bottom-up passive chart parser implemented in Prolog. GAP is applied in a cascade architecture first to *elementary discourse segments* ("clause-like units"), second to *sentential discourse segments*, and third and further to different types of *complex discourse segments* ("block", "division", "document") specified on the initial discourse segment annotation layer. Each of these segment levels is provided with its own set of *reduce rules*. Reduce rules are binary rules that describe the conditions under which two adjacent discourse segments form a new (larger) discourse segment. They are mostly derived from a discourse marker lexicon that contains combinatorial information about conjunctions and discourse adverbs (cf. Lüngen et al., 2008).

The rule component for the sentential level (where input segments are sentential discourse segments, and the top nodes of complete RST trees correspond to paragraphs of the text) was altered in six different experiments. It originally contained 73 rules derived from (the readings of) lexical discourse markers such as *beispielsweise* (indicating ELABORATION-EXAMPLE), *aber* (indicating CONTRAST), or *danach* (indicating SEQUENCE).

According to the findings discussed in Section 4, we added three rules that make reference to the CHS annotation layer in the rule component (cf. Table 6).

Table 6: Reduce rules operating on annotation layer CHS. General condition for R_0, R_1, and R_2: DS_1 and DS_2 are two adjacent discourse segments without a lexical discourse marker pointing to a relation other than subtypes of ELABORATION, and A_2 is an anaphor in the first sentence of DS_2, and A_1 is its antecedent in DS_1.

Rule	**Reduce target**	**Constraints by type of link between A_1 and A_2 on CHS**
R0	N-S, ELABORATION-DRIFT	(no further constraints)
R1	N-S, ELABORATION-CONTINUATION-OTHER	*cospec:ident* OR *cospec:paraphrase* OR *cospec:synonym* OR *cospec:addInfo*
R2	N-S, ELABORATION-DERIVATION	*bridging:setMember* OR *bridging:meronym* OR *bridging:poss*

We also introduced a ranking of rule groups and implemented the strategy that adjacent discourse segment pairs are only to be tested against reduce rules of a higher rank when no rules of a lower rank have matched before. The rule groups and their ranks are:

1. Rules based on lexical discourse markers
2. Rules based on anaphora (newly introduced)
3. Default rule (reduce target is LIST-COORDINATION, or alternatively, ELABORATION-DRIFT)

Thus, an analysis of anaphora is only activated if no discourse marker indicating a rhetorical relation other than ELABORATION and its subtypes could be found on other annotation levels.

Based on the combinations of the two versions of the default rule and the rules R_0, R_1, R_2, we conducted several parsing experiments with an article from our corpus and with the different rule sets included in GAP. The article was one that was also in the subcorpus used for deriving the statistics, as at the time of the experiments, no other articles with an annotation of anaphora was available.

Experiment I comprised the rule set of the original parser with ELABORATION-DRIFT (the most frequently occurring subtype of ELABORATION in the sample corpus) as default relation and served as a baseline. In experiment II, we tested the original rule set with LIST-COORDINATION as default relation plus the assignment of ELABORATION-DRIFT whenever any kind of anaphoric relation was found between two discourse entities in DS_1 and DS_2 (R_0 in Table 6). Experiment III comprised the original rule sets and rules R_1 and R_2 with conditions derived from the corpus study for the assignment of ELABORATION-CONTINUATION-OTHER and ELABORATION-DERIVATION according to the type of anaphora (cf. Table 6). The performance results for these discourse parsing experiments are shown in Table 7.

For deriving the figures in the column entitled "RRSet 30", the `relname` attribute in the reduce rules and in the master annotations were re-labelled by mapping all instances of subtypes of ELABORATION on one generic ELABORATION label.

Table 7: Results for discourse parsing experiments with and without anaphora processing

	Anaphora processing	**Default Relation**	**RRSet 44**		**RRSet 44**	**RRSet 30**
			Prec	**Rec**	**Rec max**	**Rec max**
I	No (Baseline)	ELABORATION-DRIFT	34.06	34.83	38.20	42.70
II	Rule R_0	LIST-COORDINATION	35.16	35.96	38.20	43.82
III	Rules R_1, R_2	LIST-COORDINATION	37.36	38.20	41.58	44.94

Using the full RRSet with 44 categories, the parser in experiment III, which included rules about subtypes of ELABORATION relations derived from specific types of anaphoric relations, performed best with a recall of 38.20% (precision 37.36%). The general assignment of the most frequent subrelation ELABORATION-DRIFT in case of an occurrence of

any kind of anaphora between DS_1 and DS_2 (experiment II) performed worse than the baseline. In experiments II and III, precision was also improved in comparison with the baseline, because rules R_1 and R_2 are more specific than the default rule of the baseline and thus filter out more hypotheses. In the third column entitled "Rec max", the maximum recall, i.e. the recall that can be reached when the whole, unpruned chart is matched against the reference file, is shown.

In the fourth column, the maximum recall for a praser with reduced relation set of 30 categories, where all subtypes of ELABORATION are represented by the general ELABORATION label is shown. The four series of experiments represented by each column all show the tendency that the performance gets better when constraints about anaphora are added (in the Rec max experiments the precision lay between 12 and 19% and showed the same tendency). However, since the increases of percentages rely on a handful of relation labels only, experiments with more documents are needed to confirm this.

6 Conclusions

Anaphoric (coreference) structure and relational, hierarchical discourse structure are two aspects of the description of coherence in discourse. In several theories of relational discourse structure, anaphora, i.e. semantic relations between discourse entities play a role in defining the ELABORATION relation. Semantic relations between (topical) discourse entities are also the basis of text structures described by thematic progression analyses. Hence we refined the original definition of ELABORATION by introducing subtypes according to different types of thematic development. In discourse analyses in the form of RST annotation of text, the ELABORATION relation was assigned to two adjacent discourse segments when no discourse markers for other standard relations like CONTRAST or SEQUENCE are available. Furthermore, we introduced a framework for the annotation of anaphora.

For an empirical investigation of the relation between discourse anaphora and discourse structure we statistically analysed a corpus that was independently annotated on the levels of anaphoric structure and rhetorical structure. The focus of the investigation has been on ELABORATION relations and whether anaphora can serve as a cue for ELABORATION, because unlike other RST relations, most subtypes of ELABORATION lack associations with lexical discourse markers. The research questions guiding our analyses were whether anaphora could be used as a necessary and/or sufficient criterion for ELABORATION, whether subtypes of ELABORATION correlate with specific subtypes of anaphora, and whether anaphora could be used as a cue in automated discourse analysis.

According to our results, anaphora is not a sufficient condition for ELABORATION, i.e. a large percentage of anaphoric instances was connected to relations other than ELABORATION. Still, anaphora seems to be a necessary condition for most subtypes of ELABORATION. The latter finding could be established after additionally annotating abstract entity anaphora in the corpus, which is frequently correlated with the subtype of

ELABORATION-DRIFT. Four ELABORATION subtypes were fairly ambiguous with respect to correlated anaphora types, but particularly ELABORATION-CONTINUATION-OTHER, ELABORATION-DERIVATION, and ELABORATION-INTEGRATION were strongly associated with *cospec:ident*, *bridging:setMember*, and *bridging:hasMember*, respectively.

The results of six discourse parsing experiments with one journal article, introducing rules operating on the CHS annotation layer in the discourse parser developed in the SemDok project, do suggest that a detailed analysis of anaphora types may help identify instances of specific subtypes of ELABORATION relations better, although the results of the test runs with a more informed evaluation of anaphora were only slightly better than those where ELABORATION was always assigned as a default relation when no other discourse marker was present.

The fact that anaphora is not a sufficient condition for ELABORATION, and the fact that ELABORATION is frequently used as a default relation could also be taken as arguments for introducing a *thematic level* as a separate and self-contained level of discourse analysis and annotation that complements RST analyses as suggested in Stede (2007). But then in order not to introduce redundancy into the representation of discourse, we think that one would also have to remove ELABORATION from the RST relation set and to relax the connectedness constraint of Mann and Thompson (1988).

References

Asher, N. (1993). *Reference to abstract objects in discourse*, volume 50 of *Studies in Linguistics and Philosophy*. Kluwer Academic Publishers, Dordrecht.

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

Bärenfänger, M., Lobin, H., Lüngen, H., and Hilbert, M. (2008). OWL ontologies as a resource for discourse parsing. *LDV-Forum. GLDV-Journal for Computational Linguistics and Language Technology*, 23(2):17–26.

Carlson, L. and Marcu, D. (2001). Discourse tagging reference manual. Technical report, Information Science Institute, Marina del Rey, CA. ISI-TR-545.

Carlson, L., Marcu, D., and Okurowski, M. E. (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, Eurospeech 2001, Denmark.

Clark, H. (1977). Bridging. In Johnson-Laird, P.N. & Wason, P., editor, *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press, Cambridge.

Corston-Oliver, S. (1998). *Computing of Representations of the Structure of Written Discourse*. PhD thesis, University of California, Santa Barbara.

Cristea, D., Ide, N., Marcu, D., and Tablan, M.-V. (2000). Discourse structure and co-reference: An empirical study. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)*, Luxembourg.

Cristea, D., Ide, N., and Romary, L. (1998). Veins theory: A model of global discourse cohesion and coherence. In *Proceedings of ACL/COLING'98*, pages 281–285, Montreal.

Daneš, F. (1970). Zur linguistischen Analyse der Textstruktur. *Folia Linguistica*, 4:72–78.

Diewald, N., Stührenberg, M., Garbar, A., and Goecke, D. (2008). Serengeti – webbasierte annotation semantischer relationen. *LDV-Forum. GLDV-Journal for Computational Linguistics and language Technology*.

Givón, T. (1983). Topic continuity in discourse: An introduction. In Givón, T., editor, *Topic Continuity in Discourse: A Quantitative Cross-Language Study*, pages 5–41. John Benjamins, Amsterdam, Philadelphia.

Givón, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics*, 30:5–55.

Goecke, D., Stührenberg, M., and Holler, A. (2007). Koreferenz, Kospezifikation und Bridging: Annotationsschema. Interne Reports der DFG-Forschergruppe 437 "Texttechnologische Informationsmodellierung".

Goecke, D., Stührenberg, M., and Witt, A. (2008). Influence of Text Type and Text Length on Anaphoric Annotation. In (ELRA), E. L. R. A., editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Harman, D. and Liberman, M. (1993). TIPSTER complete. Philadelphia: Linguistic Data Consortium.

Holler, A., Maas, J.-F., and Storrer, A. (2004). Exploiting coreference annotations for text--to-hypertext conversion. In *Proceedings of the 4th International Conference on Language Resources and evaluation (LREC 2004)*, volume II, pages 651–654, Lissabon.

Holler-Feldhaus, A. (2004). Koreferenz in Hypertexten: Anforderungen an die Annotation. *Osnabrücker Beiträge zur Sprachtheorie (OBST)*, pages 9–29.

Hovy, E. and Maier, E. (1995). Parsimonious or profligate: How many and which discourse structure relations? Unpublished paper, `http://www.isi.edu/natural-language/people/hovy/publications.html`.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic*. Kluwer, Dordrecht.

Karttunen, L. (1976). Discourse referents. *Syntax and Semantics: Notes from the Linguistic Underground*, 7:363–385.

Knott, A., Oberlander, J., O'Donnell, M., and Mellish, C. (2001). Beyond elaboration: The interaction of relations and focus in coherent text. In Sanders, T., Schilperoord, J., and Spooren, W., editors, *Text representation: Linguistic and psycholinguistic aspects*, volume 8 of *Human Cognitive Processing*, pages 181–196. Benjamins, Amsterdam.

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

Le Thanh, H., Abeysinghe, G., and Huyck, C. (2004). Generating discourse structures for written texts. In *Proceedings of COLING'04*, Geneva, Switzerland.

Lüngen, H., Bärenfänger, M., Hilbert, M., Lobin, H., and Puskàs, C. (2008). Discourse relations and document structure. In Metzing, D. and Witt, A., editors, *Linguistic modeling of information and Markup Languages. Contributions to language technology*, Text, Speech and Language Technology. Springer, Dordrecht. To appear.

Lüngen, H., Lobin, H., Bärenfänger, M., Hilbert, M., and Puskàs, C. (2006). Text parsing of a complex genre. In *Proceedings of the Conference on Electronic Publishing (ELPUB)*, pages 247–256, Bansko, Bulgaria.

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

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

O'Donnell, M. (2000). RSTTool 2.4 – A markup tool for Rhetorical Structure Theory. In *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, pages 253 – 256, Mitzpe Ramon, Israel.

Polanyi, L. (1988). A formal model of discourse structure. *Journal of Pragmatics*, 12:601–638.

Polanyi, L., van den Berg, M., and Ahn, D. (2003). Discourse structure and sentential information structure. *Journal of Logic, Language and Information*, 12:337–350.

Reitter, D. and Stede, M. (2003). Step by step: Underspecified markup in incremental rhetorical analysis. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at the EACL*, Budapest.

Stede, M. (2007). *Korpusgestützt Textanalyse. Grundzüge der Ebenen-orientierten Textlinguistik*. Gunter Narr, Tübingen.

Walsh, N. and Muellner, L. (1999). *DocBook: The Definitive Guide*. O'Reilly.

Webber, B. (1988). Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistic (ACL'88)*, pages 113–122, Buffalo. State University of New York.

Webber, B. L. (1986). So what can we talk about now? In Grosz, B. J., Sparck Jones, K., and Webber, B. L., editors, *Readings in natural language processing*, pages 395–414. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Witt, A., Lüngen, H., Goecke, D., and Sasaki, F. (2005). Unification of XML documents with concurrent markup. *Literary and Linguistic Computing*, 20(1):103–116.

Wolf, F. and Gibson, E. (2006). *Coherence in Natural Language. Data Structures and Applications*. MIT Press, Cambridge, MA.

Zifonun, G., Hoffmann, L., and Strecker, B. (1997). *Grammatik der deutschen Sprache*, volume 7 of *Schriften des Instituts für deutsche Sprache*, chapter C6 "Thematische Organisation von Text und Diskurs", pages 535–591. de Gruyter, Berlin/New York.