Underspecified discourse representation*

Markus Egg and Gisela Redeker

Abstract

This paper proposes an approach to discourse structure that builds on syntactic structure to derive that part of discourse structure that can be captured without taking recourse to deep semantic or conceptual knowledge. This contribution is typically only partial; we intend to capture this partiality in terms of *underspecified constraints* that describe (but do not enumerate) the structures a given discourse might have. This allows a rather straightforward interface from syntax to discourse and yields a clean interface to modules of discourse resolution.

1 Introduction

The analysis of discourse structure has been gaining increasing importance in Natural Language Processing. Discourse structure provides semantic information that interacts with the meaning of clauses (and other constituents of the discourse that are atoms in discourse structure) in the derivation of the full interpretation of the discourse.

Consider e.g. (1) from Asher and Lascarides (2003). In its preferred interpretation, C_2-C_5 give further details about the evening described in C_1 and C_3-C_4, about the meal described in C_2. This means that the eventualities (states of affairs) described in C_2-C_5 are part of Max's evening, and C_3-C_4 describe parts of his meal as introduced by C_2. This information goes beyond the compositionally derived semantics of C_1-C_5 and complements it. I.e., this information - as well as the result of semantic composition - is only *partial* in that it does not fully determine the interpretation of the discourse.

(1) Max experienced a lovely evening last night (C_1). He had a fantastic meal (C_2). He ate salmon (C_3). He devoured lots of cheese (C_4). He won a dancing competition (C_5).

Asher and Lascarides (2003) show that the derivation of such a fully specified discourse structure presupposes a semantic analysis of the *discourse atoms* (clauses and other constituents that are atoms in discourse structure) as well as vast amounts of conceptual knowledge. However, for any computational attempt at analysing discourse structure and its contribution to the meaning of a discourse, this raises the question of how to fulfill these presuppositions. There is as yet no system available for the computational determination of discourse-atom semantics, let alone wide-coverage representations of conceptual knowledge. I.e., modelling the interaction of

*We thank the participants of CID '05 in Dortmund and two anonymous reviewers for valuable comments.

clause- and discourse-level semantics as described in Asher and Lascarides (2003) is at present not a realistic goal for a computational approach to discourse structure.

Thus, we pursue a more modest goal in our paper, viz., to derive information on discourse structure solely on the basis of syntactic structure and an appropriate syntax-discourse interface. In this respect, we follow researchers like Marcu (1997), Schilder (2002), and Webber (2004).

Descriptions of discourse structure that are obtained in this way are characteristically only *partial*, since they use syntactic structure as the only knowledge source determining the eventual discourse structure. This suggests formalising such descriptions as *constraints* on discourse structure (Schilder: 2002), similar to the ones used in the treatment of structural ambiguity in *underspecification formalisms* (Reyle: 1993; Copestake et al.: 2005; Egg et al.: 2001).

We first introduce discourse relations and discourse structure representations and the syntax-discourse interface on which our analyses are based. For an extended example we show how an incrementally built initial underspecified discourse structure representation can be enriched by further information derivable from syntactic structure. We then defend our representations against some counterarguments raised in the literature and compare our approach to related work.

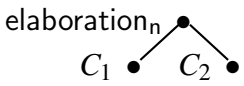
2 Discourse structure

Various systems of discourse relations have been proposed in recent approaches to the modelling of coherence and discourse connectives (Marcu: 1997; Redeker: 2000; Carlson et al.: 2003; Soricut and Marcu: 2003; Asher and Lascarides: 2003; Webber: 2004). The most explicit and elaborated one is Marcu (1997), an extension of the empirically very successful *Rhetorical Structure Theory* (RST) (Mann and Thompson: 1988). RST analyses are based on the analyst's plausibility judgments and have been applied to many text types in many languages, e.g., Dutch and German (Abelen et al.: 1993; Stede: 2004).[1]

We will base our analyses on the set of relations as defined in (classical) RST, unless otherwise stated. These relations are set in small capitals. The relations between discourse segments are sometimes (but not always) indicated by explicit discourse connectives such as *so*, *but*, or *while*. For the interpretation of Dutch connectives, we will draw on the comparative research on English and Dutch discourse connectives by Knott and Sanders (1998).

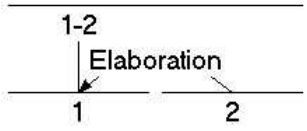
On the basis of the discourse relations, discourse structure is modelled in terms of *binary trees* in the following way: The leaves of these trees stand for the discourse atoms, while all other nodes of the tree correspond to complex discourse constituents. The label of the node for a complex constituent indicates the relation that links its immediate subconstituents. E.g., a text with two clauses C_1 (nucleus) and C_2 (satellite) related by an ELABORATION relation would schematically be depicted as follows:

[1] RST distinguishes two kinds of relations: The asymmetric *mononuclear* relations like ELABORATION or JUSTIFY relate a *nucleus* (centrally important) and a *satellite* (additional information, which could in many cases be left out without rendering the text incoherent). The symmetric *multinuclear* relations like LIST or JOINT relate discourse entities of equal status.

(2) 

Here the mother describes a functor, its daughters, the arguments of the functor. To distinguish nucleus and satellite among the arguments (where appropriate), a subscript (n or s) indicates the status of the *left* daughter.

The trees we assume for our analyses differ from those assumed in RST. Here, all the nodes are discourse units and the mother node is the convex union of all its daughter nodes. Daughters are related by different sorts of links, which also determines their position as satellite or nucleus. But our trees and RST trees have in common that all leaves are discourse atoms. (2) would be rendered as (3) in RST. The discourse structure is a tree whose mother is the segment consisting of C_1 and C_2 and whose daughters are the nucleus C_1 and the satellite C_2 elaborating on C_1:

(3) 

For the simple examples discussed so far and in the next two sections, the transformation from one type of tree to the other is straightforward, but not for nuclei with more than one satellite. We will discuss this problem in connection with the extended example (17) in section 4.2 below.

As soon as we try to account for more complex examples, we are faced with the problem that discourse structure can only be described in part. Consider e.g. (4):

(4) John is stubborn (C_1). His sister is stubborn (C_2). His parents are stubborn (C_3). So, they are continually arguing (C_4).

While C_2 and C_3 are attached to the previous discourse by implicit connectives (expressing a LIST), C_4 presents a (non-volitional) RESULT of a suitable part of the preceding discourse due to the connective *so*. This does not fix the discourse structure in (4) completely: In its preferred interpretation, C_4 is the result of C_1-C_3, i.e., due to the stubbornness of the whole family, they are constantly arguing. In another, less preferred reading, C_4 is the result of C_3 only (while John and his sister are stubborn, his parents are too, and the latter is the reason why the parents are constantly arguing), but there is no reading in which C_4 is the result of exactly C_2 and C_3.

In the following section we will show that the chosen way of representing discourse structure is capable of dealing with incomplete information on discourse structure.

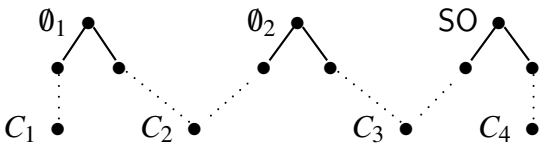
3 Representing discourse structure

This section introduces the representation of discourse structure that underlies the analyses in this paper. We describe (partial) information on discourse structure by expressions of a suitable *underspecification formalism*, here, a version of the Constraint Language for Lambda Structures (CLLS; Egg et al. 2001). First, we will present these expressions in a more intuitive way in section 3.1, then we will introduce the formal foundations for the expressions in section 3.2.

3.1 Underspecified and fully specified discourse representations

As a first example for the expressions that represent information on discourse structure, consider (5), the discourse representation for (4). Such expressions are called *constraints* and describe a number of discourse structures, which are all formalised as tree structures. The key ingredient of constraints are (reflexive, transitive and antisymmetric) *dominance relations*, which are indicated by dotted lines (see Schilder 2002 for a similar approach). Dominance of X_1 over X_2 means that X_2 is part of the structure below (and including) X_1, but there might be additional material intervening between X_1 and X_2. In these constraints and the trees they describe, '$\emptyset_n$' stands for the (very unspecific) discourse relation as introduced by the n-th implicit discourse connective,[2] SO, for the relation introduced by *so*, and C_n, for the meaning of the n-th clause of the discourse:

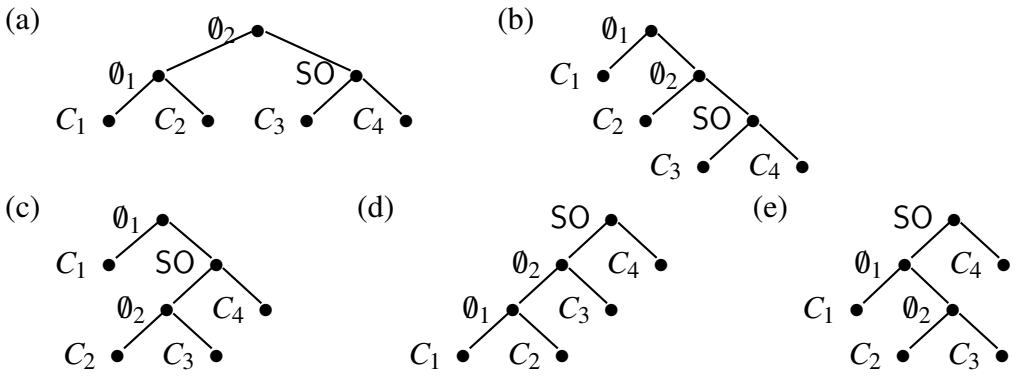
(5)



In prose: The three discourse connectives (the implicit connectives $\emptyset_1$ and $\emptyset_2$ and the explicit *so*) are all binary in that they link two text segments, which are represented as their daughters. Thus, C_1 is linked to a part of the following discourse (including at least C_2) by the implicit connective $\emptyset_1$, $\emptyset_2$ connects two discourse segments (comprising at least C_2 and C_3, respectively), and, finally, *so* connects a discourse segment to its left (which includes at least C_3) to C_4.

This constraint is compatible with a number of tree structures, called its *solutions*. If we assume that these tree structures may only comprise material that is already introduced in the constraint, then there are exactly five fully specified tree structures compatible with the constraint. These tree structures describe the potential discourse structures for (4) (see Webber 2004). (6d-e) model the preferred and (6a-b), the less preferred interpretation of (4); the inacceptable interpretation of (4) is modelled by (6c):

(6)



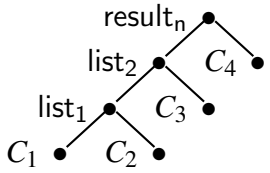
So far, we have only *arranged* the various discourse relations into a tree structure. A second task, which is crucial for the derivation of fully specified discourse structure, is the *specification* of the discourse relations. This task is due to the fact that discourse connectives themselves need

[2]Indices are merely addded to facilitate reference to different tokens of the same relation.

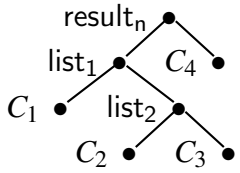
not fully determine discourse relations. For implicit discourse markers, this is quite obvious, but explicit discourse markers, too, do not always fully specify a discourse relation, which shows up e.g. in the taxonomies for English and Dutch discourse markers in Knott and Sanders (1998): In these taxonomies, the semantic contributions of explicit discourse markers are not restricted to the bottom elements, but often show up as elements higher up in the hierarchy.

For example (4) and its potential representations (6a-e), the specification of the discourse relations goes as follows. First, the semantic contributions of the implicit discourse markers (modelled as labels $\emptyset_1$ and $\emptyset_2$) are specified to a LIST relation.[3] Since lists may comprise more than two elements, we break them down into binary-branching subtrees, which add one element at a time. Second, SO, the semantic contribution of *so*, is specified to the discourse relation of non-volitional RESULT in the context of (4). (Formalisation of this step goes beyond the CLLS formalism proper, see section 3.2.2 for the technical details.) Thus, eventually, we obtain (7a) or (7b) as final representations of the preferred discourse structure of (4):

(7) (a)



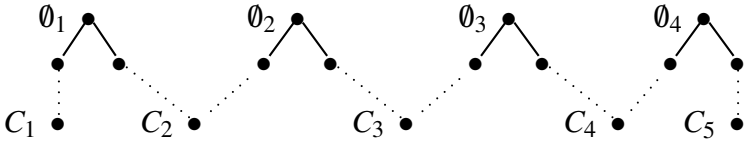
(b)



As a second example, consider the discourse structure of (8) [= (1)]. Its five sentences are connected by implicit discourse connectives, which gives rise to the constraint (9):

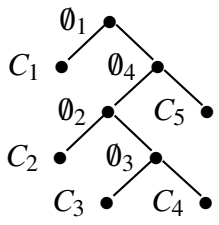
(8) Max experienced a lovely evening last night (C_1). He had a fantastic meal (C_2). He ate salmon (C_3). He devoured lots of cheese (C_4). He won a dancing competition (C_5).

(9)

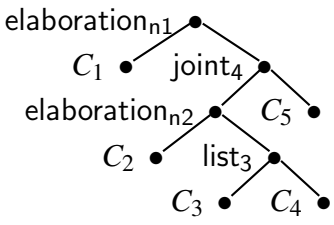


The preferred interpretation (10a) for (8) is based on one of the tree structures that are described in (9). In this tree structure, the discourse relations are not yet specified:

(10) (a)



(b)



[3] To keep track of such specifications, numeric subscripts are sometimes preserved in the tree structures. LIST models a specific kind of conjunction in RST, where the arguments must be comparable (as opposed to JOINT).

An appropriate specification models the fact that in (8), C_2-C_5 ELABORATE C_1 (the evening), C_3-C_4 ELABORATE C_2 (the meal), while C_2-C_4 (as a whole) and C_5 are related by a JOINT relation, and C_3 and C_4 form a LIST. Thus, in the eventual representation (10b) for (9), the $\emptyset$-relations are specified appropriately.

In sum, underspecified constraints on discourse structures are an efficient way of modelling partial information on discourse structure.

Before we formalise our discourse structure representations, we discuss an issue that seems to clash with our claim that discourse structures can be modelled as trees, viz., the question of what it means for a relation to link two nonatomic discourse segments D_1 and D_2. Marcu (1996) says that this is possible if and only if the relation also holds between the *nuclei* of D_1 and D_2 (if these segments are mononuclear). This condition may apply recursively. Danlos (2004, 2006) formulates this 'nuclearity principle' as follows: What looks like relations between nonatomic discourse segments are in fact relations between their nuclei, because the arguments of discourse relations can only be *discourse atoms* (and segments whose top relation is multinuclear).

But then discourse structures cannot be trees, as nodes may have several parents. Consider e.g., (10b) in Danlos' analysis: The relation joint_4 would link C_2 (the head of the segment C_2-C_4 instead of the segment as a whole) to C_5, thus, C_2 would have two parents, viz., one for the $\mathsf{elaboration}_{\mathsf{n2}}$ relation between C_2 and C_3-C_4, and one for the relation joint_4.

We regard the 'nuclearity principle' as a means of *understanding* discourse structure representations without being a part of these representations. E.g., our understanding of the tree (10b) would include the insight that, eventually, the relation joint_4 between C_2-C_4 and C_5 also means that C_2 is joined to C_5, but this claim is not hard-wired into the discourse representation. In section 5 we will make ample use of this weak version of the 'nuclearity principle'.

3.2 Formal foundations of discourse representations

After this informal introduction in the discourse representations used in this paper in the preceding section, we will now characterise them in a more formal way. We will first show how the constraints on discourse structure can be expressed in the Constraint Language for Lambda structures (Egg et al.: 2001), and how the arrangement of discourse relations into a tree structure can be handled in CLLS (section 3.2.1). For the specification of discourse relations, however, we must extend CLLS, which will be discussed in section 3.2.2.

3.2.1 Arranging discourse relations in CLLS

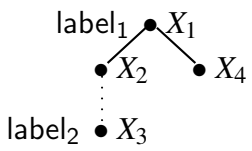
In CLLS, constraints on tree structures introduce *node variables*, *labels* for these variables, and *dominance relations* between them.[4] Intuitively, node variables correspond to discourse seg-

[4]We simplify CLLS in two ways here: First, some of the atomic constraints in CLLS (e.g., the ones for λ-binding or parallelism) are omitted. These constraints are only useful if CLLS structures are used to describe λ-*terms* as in Egg et al. (2001) (this was the original goal of the formalism, which also explains its name). Second, discourse connectives are represented as binary CLLS node labels. CLLS proper would represent them (just like the discourse atoms) as nullary labels and model the application of the connective to its arguments in terms of explicit nodes for functional application (labelled by '@'), whose daughters are nodes for functor and argument. E.g., (2) would be

ments. Labels correspond to discourse relations, they specify single discourse relations (e.g., $\mathsf{elaboration_n}$) or indicate the information of a given connective about a discourse relation (e.g., SO or $\emptyset$). For atomic segments, labels specify a unique name. Finally, dominance relations indicate those parts of a discourse structure that are not yet known.

The graphical representations for constraints used so far are shorthand for conjunctions of atomic constraints on tree structures. E.g., the constraint in (11a) is spelt out in (11b), where '$\lhd^*$' indicates dominance:

(11) (a)

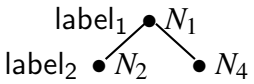


(b) $X_1 : \mathsf{label}_1(X_2, X_4) \wedge X_2 \lhd^* X_3 \wedge X_3 : \mathsf{label}_2$

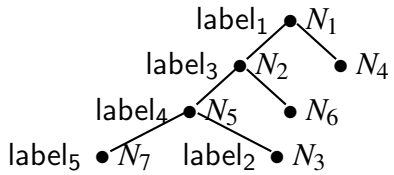
This allows us to make the intuitive partial ordering of *strength* between such constraints more precise: C_1 is at least as strong as C_2 iff C_1 comprises at least all the atomic constraints of C_2.

Now the *arrangement* of discourse relations (as given in a constraint) into a tree structure can be specified as follows: A tree structure is described by (or compatible with) a constraint, if there is a variable assignment (for the node variables of the constraint) into the domain of the tree structure (i.e., its nodes) that satisfies every atomic constraint within the constraint. E.g., both the tree structures (12a) and (12b) are compatible with (11). In (12a), both X_2 and X_3 are mapped onto N_2, which is compatible with $X_2 \lhd^* X_3$, since dominance includes identity:

(12) (a)



(b)



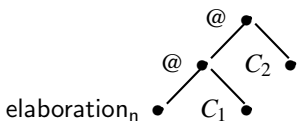
Note that actual *nodes* are distinguished from *node variables* by their names (N_n and X_n, respectively, where $n \in \boldsymbol{N}$). Graphically represented tree structures such as in (12) are also just shorthand for conjunctions of atomic relations between nodes, for instance, (12a) depicts $N_1 : \mathsf{label}_1(N_2, N_4) \wedge N_2 : \mathsf{label}_2$.

These examples illustrate that, in fact, constraints like (11) describe an infinite number of tree structures, because the material below the node assigned to X_2 is not restricted, except that it must comprise the node assigned to X_3. However, in this paper, we are only interested in so-called *constructive* solutions, where the mapping is *surjective*, i.e., every node in the solution corresponds to a node variable in the constraint.

---

rendered as (i):

(i)



3.2.2 Specifying discourse relations in an extension of CLLS

For the *specification* of discourse relations, we must extend CLLS. First, we assume a join-semilattice structure $\langle L, \leq \rangle$ for the set of labels L. Atomic elements of this structure represent the discourse relations themselves. Since the relation '$\leq$' can be interpreted as 'is more specific than', all other elements of the lattice, in particular, its greatest element $\emptyset$, model *partial information* on discourse relations. These elements comprise labels for the discourse connectives themselves (written as the name of the connective in capital letters, e.g., 'WANT' for Dutch *want* 'because'; in addition, '$\emptyset$' models the semantic contribution of the implicit discourse connective). In this way, one can represent the fact that connectives need not fully specify a discourse relation. The lattice structure formalises the intuition of Knott and Sanders (1998) that connectives can be arranged into a taxonomy.[5]

Then we can extend the above partial ordering of *strength* between constraints recursively to account for cases of different labels for the same node variable:

C_1 is at least as strong as C_2 iff C_1 comprises at least all the atomic constraints of C_2 or if all of the following conditions are met:

- $C_1 = X_n : \mathsf{label}_1(\vec{Y}) \wedge \phi_1$, where '$\vec{Y}$' stands for a specific sequence of zero or more arguments of label_1
- $C_2 = X_m : \mathsf{label}_2(\vec{Y}) \wedge \phi_2$
- $\mathsf{label}_1 \leq \mathsf{label}_2$
- ϕ_1 is at least as strong as ϕ_2

Analogously, the notion of *solution* can be extended in that an atomic constraint $X_n : \mathsf{label}_1(\vec{Y})$ can be satisfied by $N_n : \mathsf{label}_2(\vec{M})$ under a specific variable assignment iff $\mathsf{label}_2 \leq \mathsf{label}_1$ and the assignment maps X_n onto N_n, and the elements of the sequence of node variables $\vec{Y}$ onto the elements of the sequence of nodes $\vec{M}$.

Fully specified discourse structures are thus modelled as solutions of underspecified constraints on discourse structures such as (5) and (9). Solving such constraints usually involves adding further dominance relations between constraint node variables and specifying labels for discourse relations. E.g., the crucial step from (5) to a representation of the preferred reading of (4) consists in adding a dominance relation between the left daughter of the *so*- and the $\emptyset_1$-node variable. This rules out all but the last two possibilities in (6). Then the $\emptyset_1$- and $\emptyset_2$-node variables can be arranged in either order (since either connective is interpreted as a LIST relation).

In sum, the proposed approach allows a straightforward analysis of discourse structure. What is more, in this approach one can model partial descriptions of discourse structure and their resolution (or disambiguation) in terms of (monotonically) *strengthening* the involved constraints. In the next section we will show that this approach also allows a straightforward interface to syntax, i.e., a simple mapping from syntactic structure to discourse constraints.

[5]This formalisation can be used for classifications of different granularity, e.g., in the realm of conjunctive discourse relations, which is given a much finer-grained partition in Knott and Sanders (1998) than in (classical) RST.

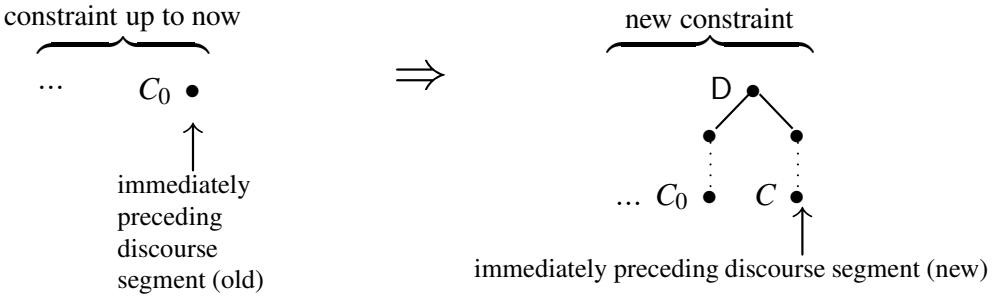
4 Constructing and resolving discourse constraints

In this section, we will first introduce the syntax-discourse interface for our analyses and then work out a larger example.

4.1 The syntax-discourse interface

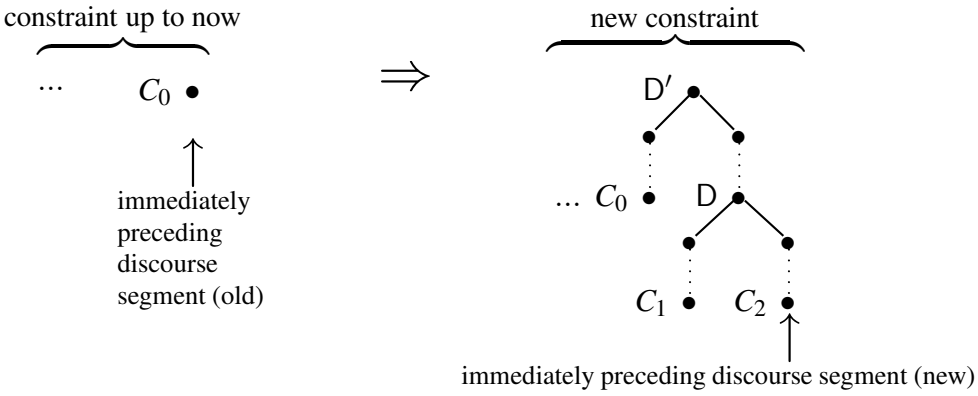
Constraints such as (5) and (9) are derived by simple interface rules: For sentences consisting of one clause C, the left daughter of the node variable that carries the label for their discourse connective dominates the node variable for the immediately preceding discourse segment C_0, its right daughter, a node variable for C. C then becomes the new immediately preceding discourse segment:

(13)



For sentences S consisting of *two* clauses C_1 and C_2 related by a discourse connective D, e.g., the fourth sentence in (17) below, the daughters of the D-node variable dominate C_1 and C_2, respectively, and the daughters of a node variable with the label for the (implicit or explicit) discourse connective D' (that links S as a whole to the preceding discourse) dominates the node variable for C_0 and the D-node variable. In addition, C_2 is determined as the new preceding discourse segment for the next sentence. (14) visualises this updating procedure:

(14)

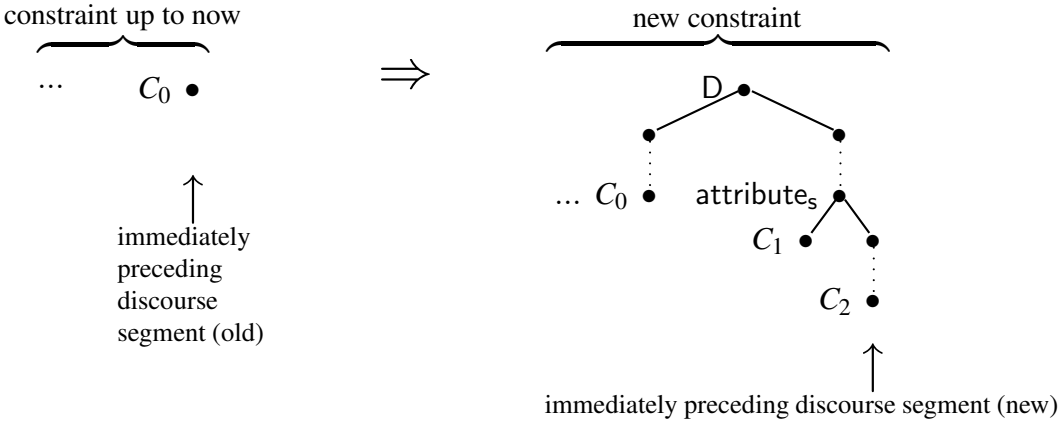


For other sentences consisting of more than one clause, additional assumptions are called for, in particular, for sentences with an *embedded* sentence S. Here S becomes the new preceding discourse segment. This is illustrated by cases such as (15), where the second sentence is linked to only the embedded sentence of its predecessor. In this way, we can model the intuition that the second sentence in (15) is also embedded in the modal context of Max's wish:

(15) Max wished that a wolf would come in. It would devour his nasty supervisor.

Formally, we can handle this low attachment of the second sentence with a rule that resembles (13), but incorporates two additional assumptions: First, we must encode the relation between matrix clause C_1 and embedded sentence C_2 in terms of a common dominating node variable of ATTRIBUTION (i.e., the wish that a wolf would come in is attributed to Max; this relation is introduced in Carlson et al. 2003). The node variable for the satellite C_1 is the left child of the ATTRIBUTION node variable, the right child of this node variable dominates the variable for C_2. Second, we determine C_2 as the immediately preceding discourse segment:

(16)



In the following section we will show how the rules (13) and (14) can be used in order to construct initial discourse constraints such as (5) and (9). These constraints can be derived incrementally, along with syntactic parsing.

However, the syntactic structure of a discourse may yield more clues to the discourse structure, which can then be used to restrain these initial descriptions of discourse structure. This two-level strategy is also employed in Schilder (2002). Such clues include the parallel structures of C_1-C_3 in (4), which strongly suggest that they should combine to form one single constituent in the discourse structure. This is bourne out by our preference for (6d) or (6e) as its discourse structure. A second clue is *modal subordination* (Roberts: 1989), which shows up in (15): The auxiliary in the second sentence indicates that this sentence is still part of the modal context introduced by the matrix verb of the first sentence. Further clues are the syntactic position of temporal clauses (Schilder: 1998) and cleft sentences (Delin and Oberlander: 1995). The extended example in the following section will illustrate such clues.

4.2 An extended example

With a larger example (from a Dutch fund-raising letter) we will now show how much information can be gathered by an appropriate syntax-discourse interface:

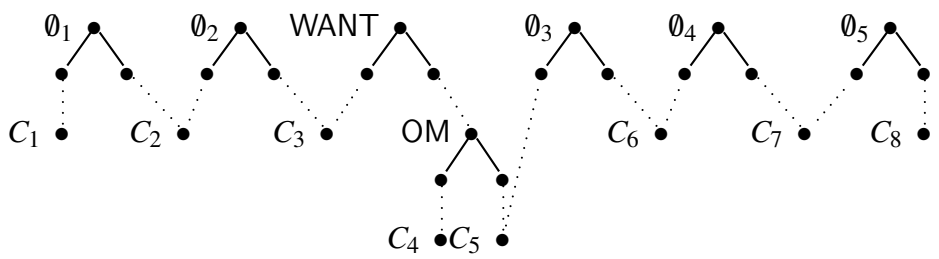
(17) Helaas raken de Nederlandse asielen iedere zomer weer boordevol met dakloze dieren. (C_1) Dieren die om welke reden dan ook door hun baasje zijn weggedaan en die nu aan hun lot zijn overgelaten. (C_2) Namens hen vragen wij om uw hulp. (C_3) Want om deze

dieren een beter bestaan te geven, (C_4) is er natuurlijk geld nodig. (C_5) Voor inentingen en sterilisaties. (C_6) Voor uitbreiding van het aantal onderkomens. (C_7) Voor extra medische zorg wanneer noodzakelijk. (C_8)

(Unfortunately, the Dutch animal shelters fill to the brim with homeless animals every summer. (C_1) Animals that have been done away with by their owner for whatever reason and that are now left to their destiny. (C_2) It is in their name that we ask your help. (C_3) Because to improve the existence of these animals (C_4) there is of course a need of money. (C_5) For vaccinations and sterilisations. (C_6) For increase of the number of shelters. (C_7) For extra medical care when necessary. (C_8))

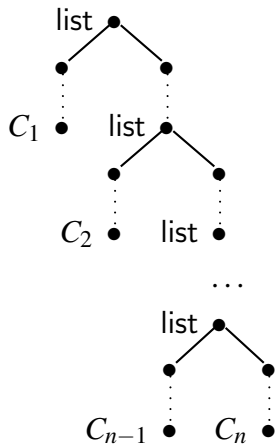
Rules (13) and (14) derive (18) as the initial discourse structure representation of (17):

(18)



To derive a fully-fledged discourse structure representation from this constraint, we take advantage of further syntactic clues, in particular, the parallel syntactic structure of C_6-C_8. We assume that these structures give rise to *lists*. We pick (arbitrarily) one of the possible ways of modelling lists in terms of binary branching, viz., (19):

(19)



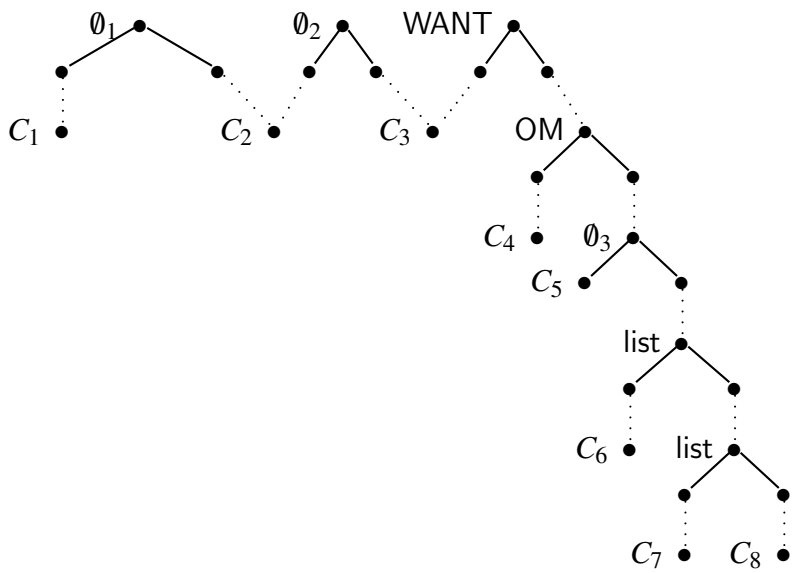
In addition, we assume that such lists as a whole are linked directly to an immediately adjacent discourse segment (if there is one). We render this linking by a discourse relation node variable (here, the one labelled $\emptyset_3$) such that its left child is the node variable for the first segment (here, C_5), and its right child dominates the node variable for the second segment.

Another potentially discourse-relevant piece of syntactic information is the fact that C_2 consists only of an NP whose head word is a direct repetition of the last word of C_1. This suggests a direct

relation between the two clauses in terms of a discourse relation node variable whose children dominate the two clauses. For (18), this is of no avail, because there already is such a node variable, viz., the one labelled $\emptyset_1$.[6]

Due to the parallel syntactic structure in C_6-C_8, constraint (18) can be strengthened to (20):

(20)



This constraint is much less ambiguous, since only the position of the $\emptyset_1$-, $\emptyset_2$- and WANT-node variables with respect to each other is not yet fixed. I.e., the ambiguity is analogous to the one in (5), there are five possible discourse structures left. Considering the fact that the number of solutions for simple 'zigzag' constraints like (5) and (9) with n discourse atoms is the Catalan number $C(n)$ of n, this considerably reduces the number of ambiguities ($C(8) = 1430$).[7] At the same time, the implicit discourse connectives $\emptyset_4$ and $\emptyset_5$ are specified to the relation list.

5 Treeness of discourse structures

This section discusses the adequacy of the proposed discourse representations. We model discourse structures by specific tree structures, where the leaves are discourse atoms and the other nodes are given by the relations between discourse constituents. These structures are more restricted than the RST-style trees (each of our trees can be mapped into an RST-style tree but not vice versa), let alone representations of discourse structures in terms of graph structures as suggested by Knott et al. (2001), Danlos (2004, 2006), or Wolf and Gibson (2005). This section will be devoted to a number of discourses that might be adduced as counterexamples proving that our notion of discourse structure is too restricted. We will show how this seemingly contrary

[6] At present, these rules have the status of hypotheses; we intend to validate them empirically against Dutch corpora like the PAROLE corpus (`http://parole.inl.nl/`) or the Corpus Gesproken Nederlands (`http://www.tst.inl.nl/cgn.htm`).

[7] One of the formulae for the Catalan number of n is $\frac{(2n)!}{n!(n+1)!}$.

evidence can be explained away. These potential counterexamples fall into three groups, which are ordered with respect to the specific problems that are claimed to emerge from the attempt to model their structure in terms of trees. These problems are *crossed dependencies*, *discontinuous constituents*, and structures with *multiple parents*.

5.1 Crossed dependencies

Wolf and Gibson (2005) claim that discourse structures often exhibit crossed dependencies, like e.g. in the following example:

(21) Schools tried to teach students history of science (C_1). At the same time they tried to teach them how to think logically and inductively (C_2). Some success has been reached on the first of these aims (C_3). However, none at all has been reached on the second (C_4).

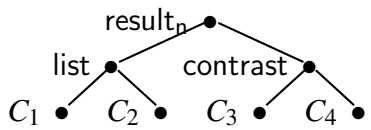
According to Wolf and Gibson (2005), C_3 links to C_1 and C_4 to C_2, as ELABORATIONS, while C_1 and C_2 on the one side and C_3 and C_4 on the other side are supposed to form CONTRASTS.

We disagree with this analysis, because it fails to take the surface structure (order and connectives) into account. The analysis can derive crossover only by assuming that the relatedness of C_1 and C_3 and the one of C_2 and C_4 should be represented as a direct relation between those segments. But the text first relates C_1 and C_2, and C_3 and C_4, respectively, and even marks those relations with connective expressions. The writer obviously gave preference to this structure over the alternative of first joining C_1 and C_3 and then C_2 and C_4.

What is more, many of the examples Wolf and Gibson (2005) adduce for cross-dependency rely on ordinary or on *complex anaphora* (Schwarz-Friesel et al.: 2004), i.e., anaphors that relate to whole sentences or larger discourse segments (abstract objects in the sense of Asher 1993). I.e., the intuition that there are dependencies in these examples that cross other dependencies can be put down to a cohesive device.

Thus, the supposed cross-dependency in (21) emerges by the complex anaphors *the first of these aims* and *the second* in C_3 and C_4, which refer back to the propositions introduced in C_1 and C_2 (that schools had the goal of teaching history of science and the goal of teaching logic and inductive thinking, respectively). Thus, the structure we would assign to (21) is (22):[8]

(22)



This example shows that discourse structure is just one possibility of organising a text. Referential anaphors can create relations between sentences that are not directly linked by discourse structure (Redeker: 1991), an additional coding of such anaphoric relations in terms of discourse structure would thus be superfluous.[9]

[8] In contrast to Wolf and Gibson (2005), we regard the relation between C_1 en C_2 as LIST and the one between C_1-C_2 and C_3-C_4 as (volitional) RESULT.

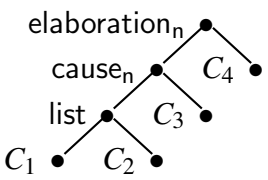
[9] Nevertheless, discourse structure and anaphoric relations are interdependent, see e.g. the results of the work in *Veins Theory* (Cristea: 2003).

An analogous explanation is available for another example adduced by Wolf and Gibson (2005) as evidence for crossed dependencies. They claim that C_4 ELABORATES C_2 only, thus crossing the relation of (non-volitional) CAUSE between C_3 and the sequence of C_1 and C_2:

(23) Susan wanted to buy some tomatoes (C_1) and she also tried to find some basil (C_2) because her recipe asked for these ingredients (C_3). The basil would probably be quite expensive at this time of the year (C_4).

We assign to (23) the structure (24), and explain the intuition that there is some dependency between C_4 and C_2 by the anaphora *the basil* in C_4, which relates back to *some basil* in C_2.

(24)



5.2 Non-continuous discourse constituents

The second group of examples that look problematic at a first glance are cases where there seems to be a non-continuous discourse constituent, which is *interrupted* by another, embedded, constituent. However, we contend that these cases do not pose a problem given our version of the 'nuclearity principle' (see section 3.1).

There are two kinds of interruptions. First, an (otherwise) atomic discourse segment is interrupted. This case shows up in (25), where *Mr. Baker's assistant for inter-American affairs, Bernard Aronson, acknowledged* is interrupted after the subject DP by another, complex discourse constituent (C_2-C_3):

(25) Mr. Baker's assistant for inter-American affairs, Bernard Aronson, (C_1) while maintaining (C_2) that the Sandinistas had also broken the cease-fire, (C_3) acknowledged: (C_4) "It's never very clear who starts what." (C_5)

This example - as well as the next one - is quoted by Wolf and Gibson (2005) from the RST Discourse Treebank (Carlson et al. 2003; from example wsj_0655).

Second, something that would in principle constitute a larger discourse segment can be interrupted at a position where one of its (potential) subconstituents ends and another one begins. This second sort of example appears typically when attributions occur within the attributed text:

(26) "The administration should now state (C_1) that (C_2) if the February election is voided by the Sandinistas (C_3) they should call for military aid," (C_4) said former Assistant Secretary of State Elliott Abrams. (C_5) "In these circumstances, I think they'd win." (C_6)

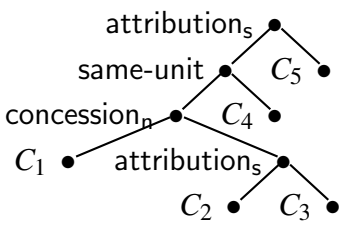
In this example, there is direct speech (C_1-C_4 and C_6), which would form a straightforward discourse constituent, were it not for the intervening ATTRIBUTION satellite C_5. In this hypothetical constituent, C_6 would ELABORATE on C_1-C_4.

We claim that these sorts of examples can indeed be assigned a tree structure, and that seeming differences between these tree structures and intuitions on the interpretation of the examples can be explained once we understand the tree structures in terms of the 'nuclearity principle'.

In addition, for the first kind of example, we need a device of indicating that two discourse segments are in fact part of one single discourse atom. Here we use the (quasi-)discourse relation SAME-UNIT as introduced by Carlson et al. (2003). It merely links the (nucleus of the) first constituent and the second constituent together.

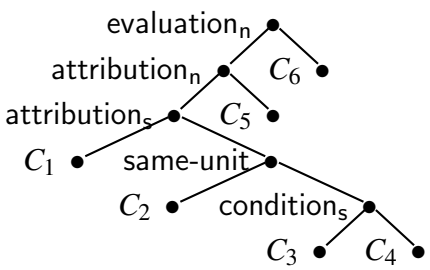
Consequently, we can uphold the analysis (27) that Carlson et al. (2002) assign to the discourse structure of (25).[10] In this structure, relating C_1-C_3 and C_4 by the relation SAME-UNIT expresses the fact that C_1 (i.e., C_1-C_3 without the satellite C_2-C_3 for the interruption) and C_4 are in principle one constituent in (25). This indicates that C_2-C_3 is a concession to C_1 and C_4 *together*:

(27)



With our version of the 'nuclearity principle' we can also account for example (26) without relinquishing the treeness of discourse structure. Its analysis (28) in the WSC Discourse Corpus (Carlson et al.: 2002) is criticised in Wolf and Gibson (2005), who claim that it fails to model two intuitions on (26): First, C_6 is part of the message linked by ATTRIBUTION to C_5, where the source is given, and, second, C_6 evaluates C_2-C_4:

(28)



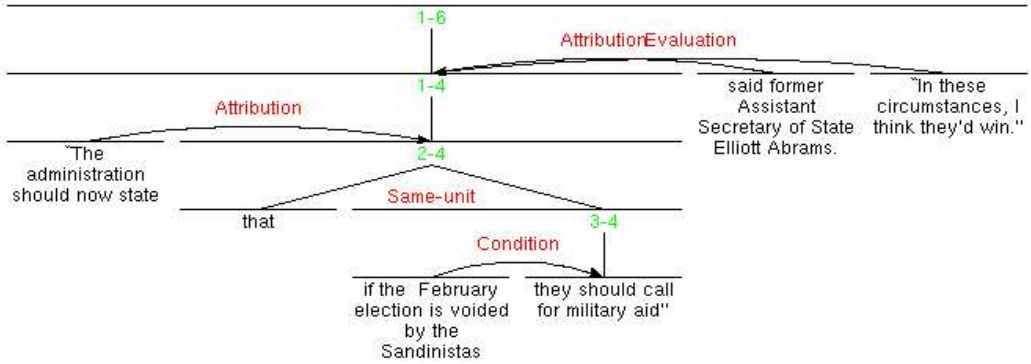
The first intuition can be reconstructed as follows: C_6 is related to C_1-C_5 by the relation of EVALUATION. Consequently, eventually, C_6 also evaluates C_2-C_4, i.e., the nucleus of the nucleus of C_1-C_5. (Since SAME-UNIT is a multinuclear relation, the iteration stops at this point.) But this intuition (that the chances of winning a military conflict under specific circumstances are evaluated) is also shared by Wolf and Gibson (2005) and supported by an instance of modal subordination: Due to its modal auxiliary, C_6 takes up the hypothetical mood of C_2-C_4.

The second intuition also follows from (28). The message attributed to the source cited in C_5 consists of C_1-C_4 and C_6: Since the source is cited in the *satellite* of the attribution relation,

[10] We deviate from their analysis in that we regard the relation between C_1 und C_2-C_3 as CONCESSION, not as ELABORATION-ADDITIONAL-E (i.e., a general elaboration relation, which interrupts another segment). Note that in an ATTRIBUTION relation, the nucleus is the message and the satellite, the segment that indicates the source.

subsequent segments can relate to the nucleus of this relation, i.e., to the message. In such cases, the message *continues* after the segment citing the source. In this example, once again, a complex anaphor (*these circumstances*) reinforces the relation between C_6 and C_2-C_4.

In the RST-style of encoding, example (26) can be modelled in an even more direct way, which straightforwardly encodes the fact that C_6 eventually evaluates C_1-C_4 (i.e., this need not be derived from the fact that C_1-C_4 is the nucleus in the constituent C_1-C_5 evaluated by C_6):



(29)

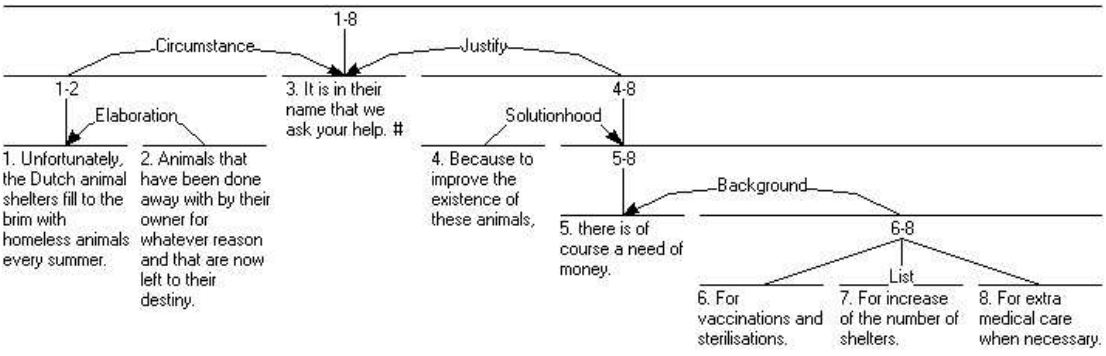
The basic idea here is that in RST-style trees, one nucleus can have several satellites. This means that an interrupting constituent such as C_5 - as long as it is a satellite (like the source in an attribution relation) - does not prevent further satellites such as C_6 from relating to the same nucleus (here, C_1-C_4).

5.3 *N*-ary RST trees

One more argument can be levied against the sort of tree structures that we use to model discourse structure in this paper. This argument upholds the claim that discourse structures are indeed trees, but only in the RST sense. However, it challenges the claim that RST trees are always binary (or can straightforwardly mapped into such trees), which means that no mapping into tree structures in our sense should be possible.

The underlying problem here has already been introduced: In RST trees, one nucleus can be associated with *several* satellites. E.g., an RST-analysis of (17) in (30) could regard C_1-C_2 as circumstance, and C_4-C_8, as justification of C_3:

(30)



Stede (2004) suggests splitting such potential *multiple-satellite constructions* (MSC) into binary parts, one for each satellite. For (17), this would be possible; then C_4-C_8 is the justification to C_1-C_3, which is internally complex (C_1-C_2 describing a circumstance of C_3).

However, for most of the potential MSC structures in fund-raising letters this would not yield plausible analyses, as it would fail to reflect the powerful rhetorical effect of symmetric justify or motivation satellites often found around the appeal to donate money (cf. Abelen et al. 1993). This may also be true for other types of persuasive texts.

But if there are discourses that can only be analysed by such MSC, a more fundamental issue emerges, viz., the question of whether the kind of tree structures we assume for discourse structure is compatible with the basic assumptions of RST: If the label of a node (or node variable) for a constituent C indicates the relation that links the immediate subconstituents of C, we cannot directly translate analyses such as (30) into a tree structure in our sense, because nodes may not have more than one label. E.g., the problem in (30) is the claim that C_3 is brought about by *two* discourse relations, which would force us to give two labels to the node for C_3.

Now the frequency of such potential MSC seems to be genre-dependent. While Carlson et al. (2003) and Stede (2004) found only few instances in their newspaper corpora, they abound in the fund-raising letters analysed by Abelen et al. (1993). The analyses on the RST website (`http://www.sfu.ca/rst/pdfs/rst-analyses-all.pdf`) corroborate this impression: In the analysed fund-raising letters (20 units altogether), five MSC are found, while all other analysed data (193 units) exhibit only 10 instances of the phenomenon. While these numbers are more illustrative than decisive, we feel that the strategy of splitting the potential MSC into binary parts, should nevertheless be empirically tested against data from a variety of genres. The representation of potential MSCs thus remains an unresolved issue, which may, however, be avoidable in non-persuasive text types.

With these comments, we conclude our analyses. They have shown that syntactic structure on its own already reveals a lot about the underlying discourse structure. In this way, one can gain valuable information that contributes to the derivation of a unique discourse structure representation for a given discourse. While we have demonstrated that a few apparent problems for the analysis of discourse structure in terms of the tree structures presented in this paper can be explained away, the question of whether all discourse structures can be modelled adequately in terms of such tree structures calls for further discussion.

6 Related work

In this section, we will compare our approach to related work. First of all, we share many intuitions with Schilder (2002). The main difference lies in the further processing of the initial underspecified discourse structure representations: Schilder uses Information Retrieval techniques (vector space model and position method) to derive full discourse structure representations from these initial representations while we merely capture the information available from syntactic structure without attempting to obtain a fully specified discourse structure representation.

The work on the Potsdam Commentary Corpus also recognises the importance of underspecification in the representation of discourse structure but implements it by chart parsing techniques

(subtree sharing and local ambiguity packing) (Stede: 2004).

The LTAG (Lexicalised Tree-Adjoining Grammar) community build their analyses of discourse structure on LTAG, which constructs syntactic tree structures for expressions from tree *fragments* associated lexically with the words in that expression. Subsequent construction of discourse structure (as well as of semantic representations) is based not on the syntax tree but on its *derivation tree*. This makes the syntax-discourse structure interface relatively complex, as can be seen e.g. in Webber's (2004) derivation of the discourse structure for (4).

What is more, potential ambiguity of a given discourse structure as e.g. for (4) must be resolved during the process of constructing it. Depending on the integration of this process into larger NLP systems, we envisage two potential problems for this strategy: If it takes place before the results of semantic construction for the discourse are available, there is the danger of not ending up with the preferred discourse structure. And if the results of semantic construction are already available, there is the question of how to let them guide the process of selecting and constructing one single discourse structure.

The proposed approach is more modular than the one based on LTAG, since it does not enforce choosing one of the possible discourse structure alternatives during discourse structure construction. This choice can be relegated to a more convenient time at which additional information (in particular, results of semantic construction) is available, which allows for a clear interface between discourse structure construction and other modules. The preferences that guide discourse structure construction on the basis of LTAG structures could be incorporated into the proposed approach as resolution preferences for underspecified discourse structure constraints.

Asher and Lascarides (2003) offer an account of discourse structures in terms of underspecified semantic representations for the involved clauses with a (possibly underspecified) discourse relation that links the respective clause to a not yet specified discourse segment. From these representations, fully specified discourse structures are built incrementally by deciding for each new clause C (a) a segment C' of the discourse structure of the previous discourse to which it attaches and (b) which discourse relation links C to C'. This is done by inference rules that use the semantics of the discourse segments involved.

The proposed constraint-based approach differs from the one of Asher and Lascarides in that we limit ourselves to *indefeasible* discourse knowledge, which is encoded in discourse structure constraints, and do not model defeasible discourse knowledge, which takes the form of inference rules. E.g., their *narration* rule states that two clauses can be related by a narration relation if they describe events that are parts of a natural event sequence, and a nonmonotonic logic infers the structure of a discourse on the basis of these rules (e.g., a defeasible Modus Ponens).

Finally, there is much common ground between our work and the work of Danlos (2004, 2006). We are both investigating the exact nature of the discourse structure and its formalisation, which involves comparing already existing approaches to discourse structure representation as well as testing potential discourse structure analyses against a wide range of data.

7 Conclusion

In this paper, we sketched an approach to discourse structure analysis. We will apply the results of this approach to discourse structure annotation. There are as yet no large-scale corpora for Dutch that are annotated for discourse structure. We are currently setting up such an annotation initiative, where we will first automatically derive partial information on discourse structure from syntactically analysed corpora. This derivation will implement the discourse-syntax interface as sketched in this paper and output discourse constraints on the basis of a suitable syntactic analysis. These constraints will then be manually specified by human annotators. Discourse annotation at the University of Potsdam (Stede: 2004) has shown that such a two-layered annotation process for discourse structure can boost inter-rater reliability and speed of corpus annotation.

Further research questions will include the search for further (indefeasible) factors to constrain discourse structure underspecification and the integration of resolution heuristics to obtain fully specified discourse structure representations. E.g., in (9), simple ontological knowledge such as the fact that salmon and cheese are meal items could be used to infer the fact that C_3 and C_4 are elaborations of C_2, which would go a long way towards resolving (9). In the future, we will also investigate the interaction between semantic construction and discourse structure construction.

References

Abelen, E., G. Redeker, and S. Thompson (1993). "The rhetorical structure of US-American and Dutch fund-raising letters." *Text 13*, 323–350.

Asher, N. (1993). *Reference to abstract objects in discourse*. Dordrecht: Kluwer.

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

Carlson, L., D. Marcu, and M. E. Okurowski (2002). RST Discourse Treebank. Corpus number LDC 2002T07, Linguistic Data Consortium, Philadelphia.

Carlson, L., D. Marcu, and M. E. Okurowski (2003). "Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory." In J. van Kuppevelt and R. Smith (eds), *Current Directions in Discourse and Dialogue*, 85–112. Dordrecht: Kluwer.

Copestake, A., D. Flickinger, C. Pollard, and I. Sag (2005). "Minimal Recursion Semantics. An introduction." *Research on Language and Computation 3*, 281–332.

Cristea, D. (2003). "The relationship between discourse structure and referentiality in Veins Theory." In W. Menzel and C. Vertan (eds), *Natural Language Processing between Linguistic Inquiry and System Engineering*. Iasi: "Al.I.Cuza" University Publishing House.

Danlos, L. (2004). "Discourse dependency structures as constrained DAGs." In M. Strube and C. Sidner (eds), *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, Cambridge, Massachusetts, USA, pp. 127–135. Association for Computational Linguistics.

Danlos, L. (2006). "Comparing RST and SDRT discourse structures through dependency graphs." This volume.

Delin, J. and J. Oberlander (1995). "Syntactic constraints on discourse structure: the case of *it*-clefts." *Linguistics 33*, 465–500.

Egg, M., A. Koller, and J. Niehren (2001). "The Constraint Language for Lambda-Structures." *Journal of Logic, Language, and Information 10*, 457–485.

Knott, A., J. Oberlander, M. O'Donnell, and C. Mellish (2001). "Beyond elaboration: The interaction of relations and focus in coherent text." In T. Sanders et al. (eds), *Text representation: linguistic and psycholinguistic aspects*, pp. 181–196. Amsterdam: Benjamins.

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

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

Marcu, D. (1996). "Building up rhetorical structure trees." In *Proceedings of the 13th National Conference on Artificial Intelligence*, Portland, pp. 1069–1074.

Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph. D. thesis, Department of Computer Science, University of Toronto.

Redeker, G. (1991). "Linguistic markers of discourse structure." *Linguistics 29*, 1139–1172.

Redeker, G. (2000). "Coherence and structure in text and discourse." In W. Black and H. Bunt (eds), *Abduction, Belief and Context in Dialogue*, pp. 233–263. Amsterdam: Benjamins.

Reyle, U. (1993). "Dealing with ambiguities by underspecification: construction, representation, and deduction." *Journal of Semantics 10*, 123–179.

Roberts, C. (1989). "Modal subordination and pronominal anaphora in discourse." *Linguistics & Philosophy 12*, 683–721.

Schilder, F. (1998). "Temporal discourse markers and the flow of events." In *Proceedings of the Workshop on discourse relations and discourse markers*, COLING/ACL 98, pp. 58–61.

Schilder, F. (2002). "Robust discourse parsing via discourse markers, topicality and position." *Natural Language Engineering 8*, 235–255.

Schwarz-Friesel, M., M. Consten, and K. Marx (2004). "Semantische und konzeptuelle Prozesse bei der Verarbeitung von Komplex-Anaphern." In I. Pohl and K.-P. Konerding (eds), *Stabilität und Flexibilität in der Semantik*, pp. 67–86. Frankfurt: Peter Lang.

Soricut, R. and D. Marcu (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of NAACL 2003*.

Stede, M. (2004). "The Potsdam Commentary Corpus." In B. Webber and D. Byron (eds), *ACL 2004 Workshop on Discourse Annotation*, Barcelona, Spain, pp. 96–102.

Webber, B. (2004). "D-LTAG: extending lexicalized TAG to discourse." *Cognitive Science 28*, 751–779.

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