Robust Text Analysis via Underspecification

Frank Schilder[*]

Department for Informatics
Vogt-Kölln-Str. 30
22527 Hamburg
Germany

Abstract

This paper is concerned with the robust analysis of the discourse structure of a text via underspecification. Most current discourse theories (e.g. Rhetorical Structure Theory (RST) by Mann and Thompson (1988), Abduction by Hobbs *et al.* (1993) or Segmented Discourse Representation Theory (SDRT) by Asher (1993)) require detailed world and context knowledge for the derivation of the discourse structure. A discourse structure for a given text has to be obtained in every case. For an ambiguous discourse a high number of structures may be generated.

The present approach instead derives an *underspecified* discourse structure for text based on a limited set of discourse cues. Only when evidence for a discourse relation or a set of discourse relations is given, for example, via a discourse marker is the discourse structure further specified.

After providing background information on underspecification and SDRT, a general framework of an underspecified discourse grammar is outlined. This framework captures scope ambiguities of discourse relations, introduces to the SDRT representation the underspecification of the discourse relation that links two segments, and further specifies the content of an abstract topic node that dominates a segment.

1. Introduction

A robust processing of text that results in the text's discourse structure is not easy to achieve. Even a small and relatively simple text presupposes an extensive body of world knowledge. The derivation of the discourse structure, however, can be useful for many text processing tasks such as automatic text summarising, text retrieval and information extraction. Hence a robust, but not too thorough analysis of a text can help to improve these tasks. So far most current discourse theories presuppose a rich knowledge representation system including an inference machine. Hence robustness is only very rarely found in discourse theories which is partly due to the complexity of the theoretical undertaking. Many questions in discourse processing are still unsolved, such as anaphora resolution.

[*] Many thanks to Christie Manning for her help and support.

A few studies that try to aim at a more robust derivation of the rhetorical structure of a text have already been carried out. Marcu (1999), for example, employs decision-based learning techniques for rhetorical parsing. A crucial prerequisite for the success of the parser, however, is a discourse corpus tagged with rhetorical and semantic information. Unfortunately, there is still a lack of such corpora and compiling these corpora is quite work intensive and time-consuming.

In addition to the need to have more robust text processing tools for Natural Language Processing tasks such as summarising, a robust and seemingly shallow modelling of discourse processing may more accurately mirror what humans do while reading a text. A reader can grasp the gist of an article even when only skimming it. On reading the same article again the reader may build a more detailed representation of the article's structure and content, but it is questionable whether she will ever build up a complete and fully specified discourse structure. In contrast, current discourse theories specify that every single segment has to be put into a hierarchical order regarding the rest of the text. There is no empirical evidence that human readers actually do such a thing and certainly they do not do it as thoroughly as current discourse theories predict. On the contrary, studies on discourse annotation, as well as psycholinguistic research, suggest that readers do not always fully specify the discourse structure and anaphoric relations within a text.

A study on discourse annotation by (Marcu *et al.*, 1999), for example, suggests that human annotators of text employ a wait-and-see-approach while tagging text according to discourse structure. Log files created during their empirical studies showed that the annotators simultaneously maintained a high number of unrelated parts of discourse. This finding contradicts the view that a newly processed discourse unit is immediately incorporated in the discourse structure derived so far. Psychological investigations also show that readers do not always specify in every detail what the rhetorical structure of a text is.[1]

Hence, the main assumption of this paper is that the discourse structure should not, and even cannot always be precisely determined. The hierarchical structure that all current discourse theories assume cannot be pinned down as concretely as these theories demand. Instead, the discourse structure is only partly spelled out. There may be some passages in the text that can be fully specified with respect to the discourse structure, but other parts of the texts may not. There the discourse structure is left *underspecified*.

In this paper, the underspecification of the discourse structure is used to develop a general framework for discourse processing. Such a framework provides a base for a system that derives a formalisation in every case, even when crucial knowledge sources are not available. Consequently, the system draws heavily on underspecification techniques as they have already been successfully employed for the semantic analysis of sentences.

So far, only a few studies have been carried out on applying underspecification formalisms to discourse grammars (Gardent & Webber, 1998; Schilder, 1998). The current paper goes beyond these approaches by focusing on the underspecified representation of all possible discourse relations. Also discussed is how the topic information covering a more concise and abstract representation of larger text spans can be incorporated into the representation. Moreover, a method for specifying an upper bound of all conceivable readings is provided before a description of a preliminary implementation and an example derivation are discussed. Finally, future extensions of the framework and the implementation are discussed in the conclusion.

[1] See Garrod and Sanford (1985) for experiments on underspecified anaphoric references.

2. Background

This section provides some background information on underspecification techniques used in sentence and discourse semantics as well as a concise introduction to SDRT. SDRT has been chosen as the basic formal framework for the given approach, because it offers a high degree of formal machinery for capturing a wide variety of discourse phenomena including discourse attachment and constraints on anaphora resolution. For a robust text analysis, however, this rich formalism is rather a hindrance. Nevertheless, it may be useful to work within such a formal framework where robust techniques on drawing inferences on world knowledge can be incorporated as soon as they are developed. The general approach taken by this paper is to leave out parts of the theory that are computationally unattractive, but allow them to be substituted by robust methods at a later time.

2.1. Underspecification

Underspecification formalisms provide a formal system that can be used for the concise representation of more than one reading for an ambiguous sentence such as (1):

(1) Every man loves a woman.

The logical form may be

a. $\forall x \exists y\ man(x) \rightarrow (woman(y) \wedge love(x, y))$, or

b. $\exists y \forall x\ man(x) \rightarrow (woman(y) \wedge love(x, y))$

Within an underspecification formalism such as the hole-semantics proposed by Bos (1995) a representation is derived that leaves open the ordering between the two quantifiers. This is done by ordering constraints between sub-formulae (i.e. $\leq$ holding for sub-formulae of the Predicate Logic). Figure 1 reflects this partial ordering between sub-formulae.

More precisely, the ordering constraints hold between labels. Note that the sub-formulae in figure 1 are labelled either as a *hole* (e.g. h_o) or a *plug* (e.g. l_1). Resolving the representation means filling the *holes* (i.e. h_0, h_1, h_2) with *plugs* (i.e. l_1, l_2, and l_3).[2]

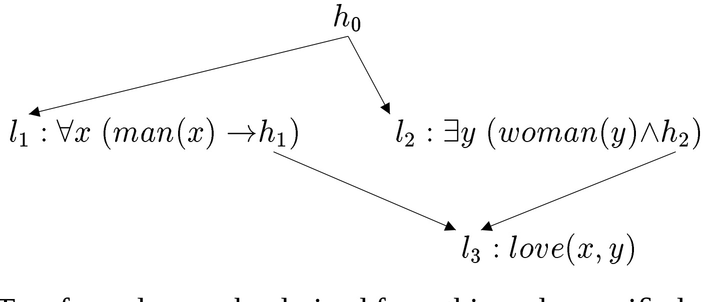
h_0
$l_1 : \forall x\ (man(x) \rightarrow h_1)$
$l_2 : \exists y\ (woman(y) \wedge h_2)$
$l_3 : love(x, y)$

Figure 1: Two formulae can be derived from this underspecified representation

Similarly, discourse grammars have been developed that also allow underspecification. Here the scope of the to-be-derived rhetorical relations may be left open. Consider (2):

[2] There are two conceivable pluggings for the underspecified representation in figure 1: (a) $h_0 = l_1 \wedge h_1 = l_2 \wedge h_2 = l_3$ and (b) $h_0 = l_2 \wedge h_1 = l_3 \wedge h_2 = l_1$.

(2) (a) I try to read a novel (b) if I feel bored or (c) I am unhappy. (Gardent & Webber, 1998)

The discourse in (2) is ambiguous with respect to the expressed discourse structure. Either the speaker tries to read a novel provided one of the two conditions in (b) and (c) hold or the speaker tries to read a book *or* she is unhappy. As Gardent and Webber (1998) show, these two readings can be represented by leaving the structural relations between scope-bearing discourse relations underspecified. A formal representation is presented in figure 2. A tree logic is used to represent several trees in one representation (i.e. forest) instead of one tree for each reading, by employing dominance constraints on node labels similar to the ordering constraints for the hole-semantics. Such constraints on node labels are imposed indicating the strict dominance relation or the dominance relation, which is transitive. The strict dominance relation (i.e. parent relation) is drawn with a straight line, whereas the dominance relation is indicated by the dotted line.[3]

The forest representation in figure 2 can give rise to the following specified readings: (i) *a if (b or c)* or (ii) *(a if b) or c*.

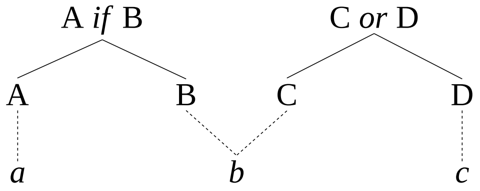

Figure 2: The underspecified discourse structure for (2).

2.2. Segmented Discourse Representation Theory (SDRT)

SDRT can be seen as a natural extension of DRT (Kamp & Reyle, 1993). Within DRT Discourse Representation Structures (DRSs) are defined as pairs $\langle U, C \rangle$, with U a (finite and possibly empty) set of discourse referents, and C a (finite) set of conditions. A shortcoming of DRT is that longer discourse are represented as a conjunction of conditions. No hierarchical structure between different discourse segments can be captured by DRT. SDRT, on the other hand, allows segmental information to be added via discourse relations. Similar to a DRS, an SDRS is defined as a pair $\langle \mathcal{U}, \mathcal{C} \rangle$, with $\mathcal{U}$ a (finite) set of discourse segments, and $\mathcal{C}$ a (finite) set of SDRS conditions. Those conditions on $\mathcal{U}$ are obtained by applying a discourse relation to the discourse segments from $\mathcal{U}$.

It is important to note that the definition of an SDRS is recursive. The universe $\mathcal{U}$ consists of discourse segments which are either DRSs (i.e. basic case) or again SDRSs. Following Asher (1996) DRSs and SDRSs will be labelled ($\{K_1, \ldots, K_n\}$). Labels will become more important for the underspecified version of SDRT. But let us first present the formal recursive definition of an SDRS given as a pair of sets containing labelled DRSs or SDRSs, and the discourse relations holding between them.

[3] The two different approaches (i.e. hole-semantics and underspecification via dominance relations) are different ways to express underspecification. Using dominance relations is a more general way to capture underspecification, since the differentiation between holes and labels is not necessary.

Definition 1 (SDRS) *Let $K_1 : \alpha_1 \ldots K_n : \alpha_n$[4] be a labelled DRSs or SDRSs and R a set of discourse relations. The tuple $\langle \mathcal{U}, \mathcal{C} \rangle$ is an SDRS if*

(a) $\mathcal{U}$ is a labelled DRS and $\mathcal{C} = \emptyset$ or

(b) $\mathcal{U} = \{K_1 \ldots, K_n\}$ and $\mathcal{C}$ is a set of SDRS conditions. An SDRS condition is a discourse relation such as $D(K_1, \ldots, K_n)$, where $D \in R$.

For the basic case (i.e. $\langle K, \emptyset \rangle$) K labels a DRS representing the semantic context of a sentence:

(3) Pedro owns a donkey.

$$K : \begin{array}{|l|}\hline x\ y\ s \\ \hline \text{Pedro}(x) \\ \text{donkey}(y) \\ \text{owns}(s,x,y) \\ \hline \end{array}$$

A clause that contains a verb constitutes a segment. For two segments a discourse relation has to be derived that furthermore introduces a hierarchical ordering indicated by a graph representation. Within this graph the nodes are the labelled SDRSs and the edges are discourse relations. Apart from the discourse relations, which impose the hierarchical ordering, 'topic' relations add more internal structure to this graph. If a sentence α is the topic of another sentence β, this is formalised as $\alpha \Downarrow \beta$.[5] This symbol also occurs in the graph, indicating a further SDRS condition. The graph representation illustrates the hierarchical structure of the discourse and in particular the open attachment site for newly processed sentences. Basically the constituents on the so-called 'right frontier' of the discourse structure are assumed to be available for further attachment (cf. Webber (1991)).[6]

As mentioned earlier, SDRT exploits discourse relations to establish a hierarchical ordering of discourse segments. How the discourse relations such as *Narration* or *Elaboration* are derived is left to an axiomatic theory called DICE (DIscourse in Commonsense Entailment) that uses a non-monotonic logic. Formally, this theory is expressed by means of the Comonsense Entailment (CE) (Asher & Morreau, 1991).

Taking the reader's world knowledge and Gricean-style pragmatic maxims into account, DICE provides a formal theory of discourse attachment. The main ingredients are defaults describing laws that encode the knowledge we have about the discourse relation and discourse processing. Two such laws are given here as an example for giving an impression of the type of information that has to be formalised:

Narration A common 'topic' is required for the two sentences α and β. It is the preferred relation for narrative texts and hence inferred by default if other information is not given.

[4] Greek symbols are normally used to described the semantic representation of sentences.

[5] A further SDRS condition is *Focus Background Pair* (FBP) which is introduced by *background*.

[6] See Asher (1996, p. 24) for a formal definition of openness in SDRT.

Elaboration The event described by the second sentence β is a part of the event of the first one α.

It may be concluded from this brief description of the theory's main ingredients that even for a rather short text an extensive body of world knowledge has to be encoded to feed the non-monotonic reasoning system. There has been some discussion within this theoretical framework to what extend this load of encoding common sense can be partly avoided. A proposal by Asher and Fernando (1997) employs underspecification. However, this proposal does not address the question of how an underspecified topic may look or how all conceivable readings can be derived for a given underspecified representation. In particular, the open attachment points for an underspecified DRS are not described.

Another extension of SDRT in Schilder (1998) gives a precise definition of the open attachment points by using a tree logic based on tree description grammar by Kallmeyer (1996). The Underspecified SDRT (USDRT), however, does not allow the underspecification of the discourse relation that links two segments, nor is the number of all conceivable readings for a given underspecified representation defined.

Both approaches lack especially a specification on how discourse markers may constrain an underspecified SDRS. Within the SDRT framework, only little work has been done on how discourse marker may constrain the derivation of the discourse structure.

3. Underspecification and discourse processing

The starting point of the current proposal to a robust discourse grammar is the underspecified version of SDRT (Asher, 1993) defined in (Schilder, 1998) called USDRT. In the following section, a further development to this theory is presented which outlines new treatments regarding (a) the underspecification of the discourse relation(s), (b) the determination of the topic within the discourse structure and (c) the derivation of the maximal number of conceivable readings for a given underspecified representation.

After formally defining an underspecified discourse structure, different ways of constraining the structure according to discourse clues are discussed in section 3.2. A short description of a partial implementation of the formalism as well as a derivation of an example discourse are given.

3.1. Underspecification via tree descriptions

The proposed formalism employs a tree logic that allows a concise representation of all conceivable discourse tree structures. Analogous to other approaches to underspecification (e.g. (Reyle, 1993; Bos, 1995; Pinkal, 1996)), the underspecification between the sub-formulae (i.e Segmented Discourse Representation Structures (SDRSs)[7]) is expressed by labels and the (immediate) dominance relations that hold between these labels specify the ordering between daughter nodes. The definition of an underspecified USDRS is as follows (cf. (Schilder, 1998)):

[7] The semantic content of a sentence is represented by a DRS, larger sequences by an SDRS.

Definition 2 (USDRS) *Let S be a set of DRSs, L a set of labels, $\mathcal{R}$ a set of discourse relations. Then U is a USDRS confined to the tuple $\langle S, L, \mathcal{R}\rangle$ where U is a finite set consisting of conditions of the following form:*

- *structural information*
  - *immediate dominance relation: $K_1 \lhd K_2$, where $K_1, K_2 \in L$*
  - *dominance relation: $K_1 \lhd^* K_2$, where $K_1, K_2 \in L$*
  - *precedence relation: $K_1 \prec K_2$, where $K_1, K_2 \in L$*
  - *equivalence relation: $K_1 \approx K_2$, where $K_1, K_2 \in L$*
- *content information*
  - *sentential (i.e. universe): $s_1 : \alpha$, where $s_1 \in L, \alpha \in S$*
  - *segmental (i.e. conditions):*
    - *discourse relation(s) connecting two segments: $K_{R1} : relation(\mathcal{P}, K'_{R1}, K''_{R2})$, where $\mathcal{P} \subseteq \mathcal{R}$, and K_{R1}, K''_{R1}, and $K''_{R1} \in L$*
    - *topic information: $K^{\mathcal{T}}_{R1} : \mathcal{T} \subseteq \{\alpha, \beta\}$*

A USDRS consists on the one hand of content information specifying the DRSs and the conditions imposed on them. In contrast to the original definition of USDRT, a discourse relation *set* $\mathcal{P}$ provides the link between (S)DRSs. Former approaches to underspecification of discourse structure (Asher & Fernando, 1997; Schilder, 1998) do not provide an appropriate formalisation for the underspecification of the discourse relations. These approaches deal with underspecified discourse relations in the same way as scope ambiguity. However, there is a crucial difference between these two forms of ambiguity: scope ambiguity can easily be resolved by computing all combinations of scope-bearing operators. The discourse relations, on the other hand, cannot be resolved by determining the scope of all relations. The relations have to be inferred from world knowledge and the information provided by the context.

Within the standard SDRT account, only one relation must be obtained by considering world knowledge as well as additional discourse knowledge. Applying this system leads to a disambiguation of the given discourse with all conceivable readings. The SDRT approach is problematic with respect to the following two aspects. Firstly, the non-monotonic reasoning system comes with computational costs that one may not want to bear. Secondly, deriving all readings for an ambiguous discourse could be computationally intractable, since all conceivable readings are derived. Hence, any derivation within the modified version of USDRT presented here starts with a structure as shown in figure 3.[8]

In the following, important features of this underspecified representation are described in more detail.

3.1.1. Underspecified discourse relations.

In the case that the discourse relation is not known for two segments then $\mathcal{P} = \mathcal{R}$. After taking into account further restrictions, only a subset of discourse relations is possible. The underspecification of the discourse relation set $\mathcal{R}$ is expressed via a lattice structure. The set

[8] The description for the tree is $K_\top \lhd^* K^{\mathcal{T}}_{R1} \wedge K^{\mathcal{T}}_{R1} \lhd K_{R1} \wedge K_{R1} \lhd K'_{R1} \wedge K_{R1} \lhd K''_{R1} \wedge K'_{R1} \lhd^* s_1 \wedge K''_{R1} \lhd^* s_2$.

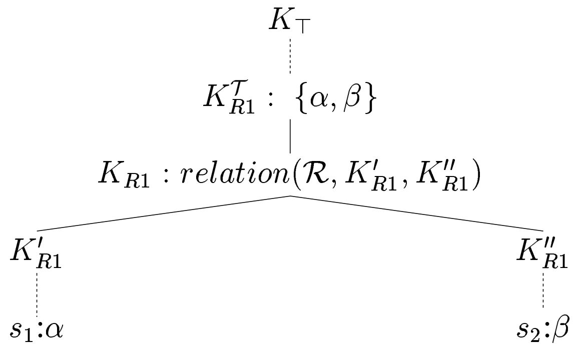
$K_\top$
$K^{\mathcal{T}}_{R1} : \{\alpha, \beta\}$
$K_{R1} : relation(\mathcal{R}, K'_{R1}, K''_{R1})$
K'_{R1}
K''_{R1}
$s_1{:}\alpha$
$s_2{:}\beta$

Figure 3: Underspecified discourse structure

of the relations can therefore easily be constrained by means of an intersection operation. The ordering of the discourse relation lattice for the four relations *Narration, Result, Elaboration* and *Explanation*, for example, can be found in figure 4.

There has been some discussion of how many discourse relations there are. The number of relations proposed by different approaches to discourse range from two (Polanyi, 1988) to as many as needed (Mann & Thompson, 1988). Still, the successful application of the current proposal does not depend on the outcome of this discussion. For an actual implementation only a subset of relations may be chosen including, for instance, *Explanation*. The output of such a system would miss many relations and dependencies expressed by the text, but still be able to cover at least all causal relations that hold between the described situations.[9]

For the theoretical considerations and the constraints on anaphora resolution two relation sets are particularly important: the subordinating relation set $\overline{\mathcal{S}}$ (e.g. *Narration*) and the subordinated relation set $\underline{\mathcal{S}}$ (e.g. *Elaboration*). A relation from the latter set allows attachment to both discourse segments, whereas the former set consists of relations that close off the preceding discourse.

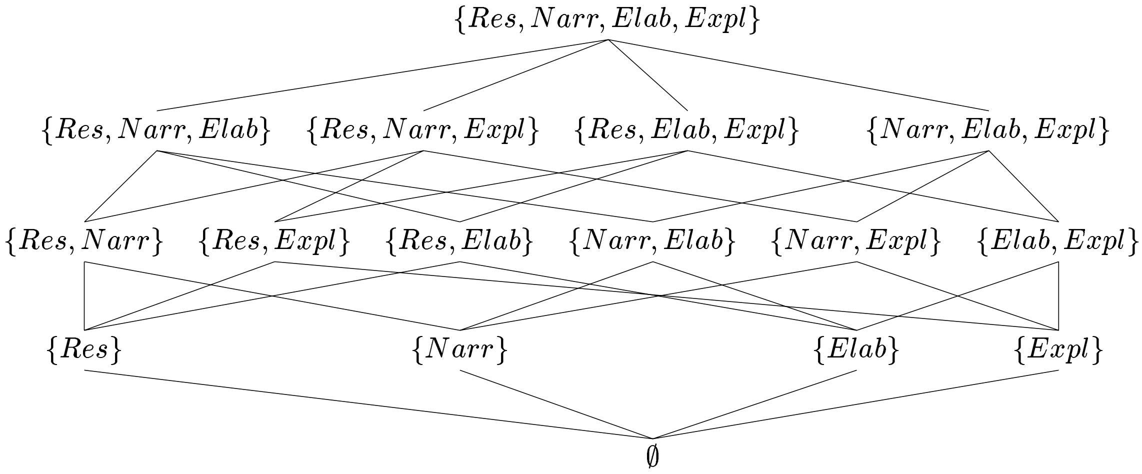

Figure 4: The discourse relation lattice for four discourse relations

[9] The number of used discourse relations is currently being further investigated. A starting point for this investigation is the work on discourse clues by Knott (1996).

3.1.2. Topic information

The topic node plays an important role for the discourse representation. It contains information in an abstract form as to what the given segment is about. Note that the usage of the term topic has varied widely in the literature. Some researchers (e.g. (Sgall *et al.*, 1986; Büring, 1997)) understand topic as a part of a sentence indicated by the linguistic surface structure. The topic structure to be investigated by the current paper, however, goes beyond the surface structure and covers larger text spans.

The definition of a common topic (i.e. $\Downarrow$) was already introduced by standard SDRT, but only as a further restriction regarding discourse attachment. I adopt the topic node as defined in Schilder (1998). In this case, the topic node is given as an additional feature for every segment.[10] However, it is not entirely clear what this node contains. For a first approximation on the content of the topic node, two types of discourse structures for two segments α and β are distinguished:

1. a subordinating structure is triggered by discourse relations such as *Narration* or *Result*. These relations close off the preceding discourse. Consequently, only the last mentioned segment β is accessible for the following discourse and the topic node gets filled by it.

2. a subordinated structure is derived for discourse relations such as *Elaboration* or *Explanation*. Here both segments remain open for attachment, the topic node gets filled by the first segment α.

Additionally, I allow a third (preliminary) discourse structure that also has an effect on the topic node:

3. a coordinated structure does not distinguish between the two segments. Both segments end up in the topic node (cf. figure 3).

The question of how to specify the topic node is the subject of other current research. For the time being, the node can contain these three types of sets reflecting (1) a subordinating, (2) a subordinated or (3) a coordinated structure. The last structure is also applied when the discourse structure is left underspecified.

3.1.3. Derivation of all readings

Note that for an underspecified representation the number of conceivable readings grows quite rapidly. Ten clauses connected via nine discourse relations have 4862 different discourse tree structures. The number of all conceivable readings can be computed via the Catalan number:[11] $C_n = \frac{(2n)!}{(n+1)!n!}$. The Catalan number provides the solution for an extensive body of combinatorial problems. The number C_n describes, for instance, the maximal number of rooted binary trees with n internal nodes. Binary trees are also the representation of the discourse structure described by USDRT with the exemption of having an additional internal topic node. Note that the discourse relation(s) always relate *two* segments. Hence the number of possible discourse structures for *n* discourse relations is C_n.

[10] The value of the topic node can be compared to the nucleus in RST.

[11] See Sloane, N. J. A. Sequences A000108/M1459 in "An On-Line Version of the Encyclopedia of Integer Sequences." `http://www.research.att.com/~njas/sequences/eisonline.html`

In addition to the discourse *tree* structure, USDRT also determines the content of the node. For a fully specified discourse structure, one specific discourse relation can be derived. The number of possible readings therefore depends on the number of discourse relations. Consequently, the total number for a discourse structure containing of n discourse relations sets (i.e. internal nodes) is $C_n \times \mid \mathcal{R} \mid^n$.

Proof 1 (sketch) The maximal number of discourse structures for a USDRS with $n+1$ segments connected via n discourse relations is determined by the Catalan number C_n. Assuming that $\mathcal{R}$ is the set of all conceivable discourse relations, there are at most $C_n \times \mid \mathcal{R} \mid^n$ different discourse trees.

We show via induction that every underspecified discourse structure of n clauses can be translated into a rooted binary tree with n leaves. Remember that the Catalan number gives the number of possible trees for a given rooted binary tree with $n+1$ leaves:

The top node $\top$ is the root of the given tree. SDRSs as defined in Definition 1 are binary tree structures, because the discourse relation possesses only two arguments. Consequently, we obtain C_n different discourse structures for $n+1$ clauses.

Finally, it has to be shown that the $relation$ node can vary with respect to the derived discourse relation. For $\mid \mathcal{R} \mid$ possible discourse relations, there are $\mid \mathcal{R} \mid^n$ different ways of assigning a discourse relations to the n internal $relation$ nodes: by assigning a unique number out of $\{1, \ldots, \mid \mathcal{R} \mid\}$ to every relation in $\mathcal{R}$, the internal nodes of the discourse structure can be represented as a $\mathcal{R}$-nary number. There are $\mid \mathcal{R} \mid^n$ different numbers for a given $\mathcal{R}$-nary number of length n.

3.2. Constraining the underspecified representation

There are several steps for determining a more specified representation of the discourse structure. First, all discourse units have to be extracted. Discourse units are clauses that contain a verbal phrase or are separated by punctuation.[12] Second, the discourse structure is built. This can be done with different degrees of specification.

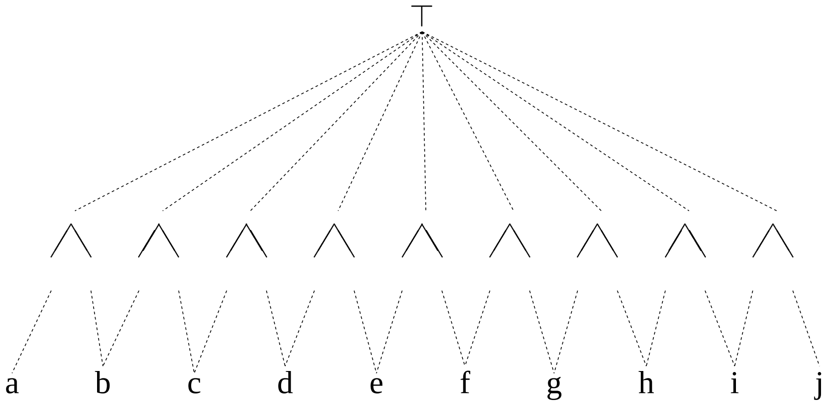

Figure 5: the discourse structure underspecified

[12] The definition of a discourse unit varies among discourse theories. More research is needed here to determine precisely what a discourse unit constitutes. The current definition is only a working hypothesis.

1. **Total underspecification.** Discourse units are connected via the set of all possible rhetorical relation $\mathcal{R}$. However, no restriction is given here and hence the number of readings that is covered by the representation grows according to the Catalan number. The underspecified tree structure figure 5 allows for $4862\times \mid \mathcal{R} \mid^9$ different readings.

2. **Underspecification restricted by discourse markers.** The totally underspecified representation can be further restricted when discourse clues are taken into account.

3. **(Partial) resolution via world knowledge.** Finally, the discourse structure can be resolved, or partially resolved, or to different degrees restricted, provided the appropriate world and background knowledge is available.[13] How far the structural ambiguities can be eliminated depends on how well the encoded theory covers world and background knowledge.

Clearly, for a robust processing only total underspecification or underspecification restricted by discourse markers are available. However, future developments on robust processing of world knowledge can easily be incorporated.

3.2.1. Implementation

The proposed discourse theory has been partly implemented. First, a discourse grammar taking into account punctuation and discourse clues determines the discourse units and has as an output a USDRS.[14] In the following some rules are named in Definite Clause Grammar (DCG) notation.

```
%% a discourse grammar fragment (without discourse semantics)
%% in DCG notation
%%
%% a discourse may be a sentence or a question.
d(P2) --> s(P2).
d(P2) --> q(P2).

%% a discourse consists of
%% a discourse clue <discourse marker|empty>,
%% a clause,
%% a discourse clue <discourse marker|(punctuation, discourse marker)|empty>,
%% and  another disourse
d --> dclue(D), cl, D, d.

% a sentence
s --> cl, fullstop.

% a question
q --> cl, questionmark.

fullstop --> ['.'].
questionmark --> ['?'].
```

[13] Certain types of domain-specific knowledge would be fairly easy to formalise.

[14] The current implementation, however, does not consider all clues that could constrain the discourse structure. Those clues are to be determined on the result of an extensive corpus study.

```
%% a clause consists of words
cl --> words.
words --> word, words.
words --> [].

%% a word must not be a punctuation sign or a discourse marker
word --> [W],{<not a punctation sign or a discourse marker>}.

% lexicon look up
dclue(D)--> {lexicon(dclue,Word,D)}, Word.

%% lexicon
lexicon(dclue,[’Contrary’, to],[’,’]).
lexicon(dclue,[],[’.’, ’Yet’]).
...
```

The discourse semantics is derived during the parse of the discourse. The tree descriptions are encoded in the following way:

```
td(<Holes>, <Trees>, <Dominance>)
```

`<Holes>` is a list that contains the set of labels that dominate other labels (e.g. h in $l \lhd^* h$) and can be “plugged” by an appropriate other label. `<Trees>` is a list of fully specified trees presented in the following general form:

```
<Mothernode>/[<Daughter1>,...,<DaughterN>].
```

Remember that the nodes are also labelled (e.g. `T1:Topic/R1:relation(R)/[K1,K2]`). And finally the dominance constraints for the tree description can be found in `<Dominance>` (e.g. `leq(K1,T2)`). Given this representation, all conceivable readings are calculated by using Bos’ plugging algorithm (Bos, 1995).[15]

3.2.2. Underspecified derivation

Let us now go through an example text and derive an underspecified representation for the given discourse structure.

(4) (a) CONTRARY to some headlines at the end of last week, (b) America’s stock-market bubble has not burst. (c) Yet the market turmoil has prompted one topical economic question: (d) how much might a crash hurt America’s economy?

(e) The answer of many American optimists is that (f) a slump in share prices would not trigger a recession, (g) because the real economy is fundamentally so sound. (h) *It* is, (i) *they* argue, much healthier than Japan’s in the late 1980s or East Asia’s economies in the mid-1990s, just before their bubbles burst. (j) It is certainly true that America has much to boast about: [. . .] (source: *The Economist*)

[15] A more efficient algorithm such as recently proposed by (Koller *et al.*, 2000) could easily be adopted for the implementation.

The totally underspecified representation for the given text can be found in figure 5. There are ten clauses connected via nine discourse relation sets. This rather short segment allows already for $4862 \times \mid \mathcal{R} \mid^9$ different discourse structures.

To restrict this number, discourse clues and punctuation signs are taken into account. In sequence (4) the discourse clues are underlined. The first clause contains the marker *contrary to*. This discourse cue phrase expresses a contrast.[16] Hence the discourse relation *Contrast* is derived for the relation set connecting the first two clauses. According to the constraints on openness defined by SDRT, this relation closes off preceding discourse segments. Consequently, the second clause (4b) ends up in the topic node.

The next clause (4c) also contains a discourse cue that expresses a contrast (i.e. *Yet*). Again the discourse relation *Contrast* can be derived.

The next clue we can get comes from the punctuation. The double column indicates an explanation in the given context. However, there are other contexts where the double column triggers a direct speech instead. Since other indicators (e.g. quotes) are missing, the relation *Explanation* can be determined for (4c+d).

For the following clause (4e) no discourse cue can be found. Accordingly, the set of all conceivable discourse relations $\mathcal{R}$ is assigned to connect (4d) and (4e).

The entire sentence (4e-f), that consists of three clauses, exhibits the same scope ambiguity as already analysed by Gardent and Webber for example sequence (2). Note that although the discourse cue *because* triggers an *Explanation* relation, the attachment site is underspecified (see figure 6).

After considering the discourse cues in (4a-g), the resulting underspecified discourse structure represents $4 \times \mid \mathcal{R} \mid^6$ different readings. Originally, this part of the example sequence could have had $132 \times \mid \mathcal{R} \mid^6$ different discourse structures.

Finally, I would like to highlight the influence of the discourse structure on the set of possible antecedents for anaphoric expressions. Note that the conceivable antecedents for the pronouns *it* and *they* in (4h) and (4i) are still accessible (i.e. *America's economy* in (4d)/*the real economy* in (4g) and *American optimists* in (4e)).

4. Conclusions and further directions

The current paper has shown how an underspecified representation of discourse structure can provide a robust representation format for text analysis. A text is first analysed as an underspecified discourse structure of $n + 1$ clauses connected by n discourse relation sets. It was also shown that the number of possible readings can be computed by the Catalan number C_n. The totally underspecified representation can furthermore be further restricted by the discourse cues found in the text.

Summarising, an underspecified version of SDRT (Schilder, 1998) was extended and the following features were added:

- Underspecifying the conceivable discourse relations via a lattice structure

[16] Considering Knott's taxonomy there are several kinds of coherence relations expressing a contrast. For the time being only a very general *Contrast* relation is assumed following the SDRT account that does not make a finer distinction.

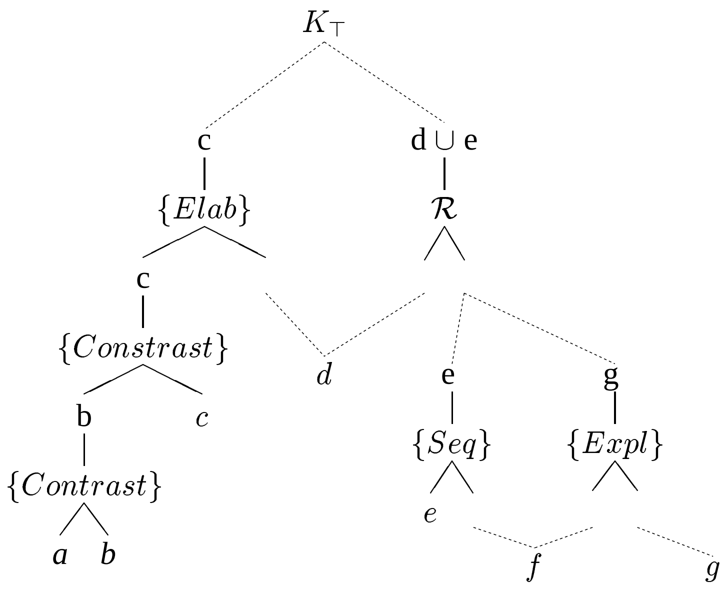

Figure 6: the discourse structure partially resolved according to discourse markers

- Restricting the set of possible readings by discourse cues

Directions of current and future work are:

- The contribution of cue phrases, especially punctuation and formatting cues (e.g. section, paragraph formatting)
- The relationship between the underspecified discourse structure and the set of possible antecedents for anaphoric expressions
- The determination of the topic for a given text segment

References

ASHER N. (1993). *Reference to abstract Objects in Discourse*, volume 50 of *Studies in Linguistics and Philosophy*. Dordrecht: Kluwer Academic Publishers.

ASHER N. (1996). Mathematical treatments of discourse contexts. In P. DEKKER & M. STOKHOF, Eds., *Proceedings of the Tenth Amsterdam Colloquium*, p. 21–40: ILLC/Department of Philosophy, University of Amsterdam.

ASHER N. & FERNANDO T. (1997). Labeling representations for effective disambiguation. In *Proceedings of the 2^{nd} International Workshop on Computational Semantics (IWCS-II)*, p. 1–14, Tilburg, The Netherlands.

ASHER N. & MORREAU M. (1991). What some generic sentences mean. In H. KAMP, Ed., *Default Logics for Linguistic Analysis*, number R.2.5.B in DYANA Deliverable, p. 5–32. Edinburgh, Scotland: Centre for Cognitive Science.

BOS J. (1995). Predicate logic unplugged. In P. DEKKER & M. STOKHOF, Eds., *Proceedings of the ninth Amsterdam Colloquium*: ILLC/Department of Philosophy, University of Amsterdam.

BÜRING D. (1997). *The Meaning of Topic and Focus - The 59th Street Bridge Accent*. London: Routledge.

GARDENT C. & WEBBER B. (1998). Describing discourse semantics. In *Proceedings of the 4th TAG+ workshop*, Philadelphia, USA.

GARROD S. C. & SANFORD A. J. (1985). On the real-time character of interpretation during reading. *Language and Cognitive Processes*, **1**, 43–61.

HOBBS J., STICKEL M., APPELT D. & MARTIN P. (1993). Interpretation as abduction. *Artificial Intelligence*, **63**(1-2), 69–142.

KALLMEYER L. (1996). *Underspecification in Tree Description Grammars*. Arbeitspapiere des Sonderforschungsbereichs 340 81, University of Tübingen, Tübingen.

KAMP H. & REYLE U. (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language*, volume 42 of *Studies in Linguistics and Philosophy*. Dordrecht: Kluwer Academic Publishers.

KNOTT A. (1996). *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, Department of Artificial Intelligence, University of Edinburgh.

KOLLER A., MEHLHORN K. & NIEHREN J. (2000). A polynomial-time fragment of dominance constraints. In *Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics*, Hong Kong.

MANN W. & THOMPSON S. (1988). Rhetorical structure theory: Toward a functional theory of text organisationn. *Text*, **8**(3), 243–281.

MARCU D. (1999). A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the ACL*, p. 365–372, Maryland, MD.

MARCU D., ROMERA M. & AMORRORTU E. (1999). Experiments in constructing a corpus of discourse trees: Problems, annotation choices, issues. In *Proceedings of the Workshop on Levels of Representation in Discourse*, Edinburgh, Scotland, U.K.

PINKAL M. (1996). Radical underspecification. In P. DEKKER & M. STOKHOF, Eds., *Proceedings of the Tenth Amsterdam Colloquium*, p. 587–606: ILLC/Department of Philosophy, University of Amsterdam.

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

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

SCHILDER F. (1998). An underspecified segmented discourse representation theory (USDRT). In *Proceedings of the 17th International Conference on Computational Linguistics (COLING '98) and of the 36th Annual Meeting of the Association for Computational Linguistics (ACL '98)*, p. 1188–1192, Université de Montréal, Montréal, Québec, Canada.

SGALL P., HAJIČOVÁ E. & PANEVOVÁ J. (1986). *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Dordrecht: Reidel.

WEBBER B. L. (1991). Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, **6**(2), 107–135.