Expressivity and comparison of models of discourse structure

Antoine Venant, Nicholas Asher, Philippe Muller, Pascal Denis, Stergos Afantenos

▶ **To cite this version:**

Antoine Venant, Nicholas Asher, Philippe Muller, Pascal Denis, Stergos Afantenos. Expressivity and comparison of models of discourse structure. SIGDIAL 2013 - Special Interest Group on Discourse and Dialogue Conference, Aug 2013, Metz, France. pp.2–11. hal-00838260

HAL Id: hal-00838260
https://inria.hal.science/hal-00838260

Submitted on 24 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Expressivity and comparison of models of discourse structure

Antoine Venant[1] **Nicholas Asher**[2] **Philippe Muller**[1] **Pascal Denis**[3] **Stergos Afantenos**[1]
(1) IRIT, Toulouse University, France, (2) IRIT, CNRS, France (3) Mostrare, INRIA, France *

Abstract

Several discourse annotated corpora now exist for NLP. But they use different, not easily comparable annotation schemes: are the structures these schemes describe incompatible, incomparable, or do they share interpretations? In this paper, we relate three types of discourse annotation used in corpora or discourse parsing: (i) RST, (ii) SDRT, and (iii) dependency tree structures. We offer a common language in which their structures can be defined and furnished a range of interpretations. We define translations between RST and DT preserving these interpretations, and introduce a similarity measure for discourse representations in these frameworks. This will enable researchers to exploit different types of discourse annotated data for automated tasks.

1 Introduction

Computer scientists and linguists now largely agree that representing discourse structure as a hierarchical relational structure over discourse units linked by discourse relations is appropriate to account for a variety of interpretative tasks. There is also some agreement over the taxonomy of discourse relations —almost all current theories include expressions that refer to relations like *Elaboration, Explanation, Result, Narration, Contrast, Attribution*. Sanders, Spooren, and Noordman 1992; Bateman and Rondhuis 1997 discuss correspondences between different taxonomies.

Different theories, however, assume different sets of constraints that govern these representations; some advocate trees: RST Mann and Thompson 1987, DLTAG Webber et al. 1999; others, graphs of different sorts: SDRT Asher and Lascarides 2003, Graphbank Wolf and Gibson 2005. Consider:

(1) [“he was a very aggressive firefighter.]$_{C_1}$ [he loved the work he was in,”]$_{C_2}$ [said acting fire chief Lary Garcia.]$_{C_3}$. [”He couldn’t be bested in terms of his willingness and his ability to do something to help you survive.”]$_{C_4}$ (from Egg and Redeker 2010)

Using RST, Egg and Redeker 2010 provide the tree annotated with nuclearity features for this example (given by the linear encoding in (s_1)), while SDRT provides a different kind of structure (s_2). Dependency trees (DTs), similar to syntactic dependency trees and used in Muller et al. 2012 for automated parsing, give yet another representation (s_3). Elab stands for elaboration, Attr for attribution, and Cont for continuation.

$$Elab_1(Attr(Elab_2(C_{1N}, C_{2S})_N, C_{3S})_N, C_{4S}) \quad (s_1)$$

$$Attr(\pi, C_3) \wedge \pi : \mathrm{Elab}(C_1, \pi_1) \wedge \pi_1 : \mathrm{Cont}(C_2, C_4) \quad (s_2)$$

$$Elab_1(C_1, C_2) \wedge \mathrm{Attr}(C_1, C_3) \wedge \mathrm{Elab}(C_1, C_4) \quad (s_3)$$

Several corpora now exist annotated with such structures: RSTTB Carlson, Marcu, and Okurowski 2002, Discor Baldridge, Asher, and Hunter 2007, GraphBank[1]. But how exactly do these annotations compare? In the illustrative example chosen and for the relation types they agree on (Elaboration and Attribution), different annotation models and theoretical frameworks invoke different numbers of instances of these relations and assign the instances different arguments or different scopes, at least on the surface. In this paper we develop a method of comparing the scopes of relations in different types of structures by developing a notion of *interpretation* shared between different structures. This interpretation specifies the set of *possible scopes of relations* compatible with a given structure. This theoretical work is important for furthering empirical research on discourse. Discourse annotations are expensive. It behooves researchers to use as much data as they can, annotated in several formalisms, while pursuing prediction or evaluation in their chosen theory. This paper provides a theoretical basis to do this.

What a given structure expresses exactly is often not clear; some discourse theories are not completely formalized or lack a worked out semantics. Nevertheless, in all of them rhetorical relations have semantic consequences bearing on tasks like text summarization, textual entailment, anaphora resolution, as well as the temporal, spatial and thematic organization of a text Hobbs, Stickel, and Martin 1993; Kehler 2002; Asher 1993; Lascarides and Asher 1993; Hobbs, Stickel, and Martin 1993; Hitzeman, Moens, and Grover 1995, *inter alia*. Theories like SDRT or Polanyi et al. 2004 adopt a conception of discourse structure as logical form. Discourse structures are like logical formulae and relations

*This research was supported by ERC grant 269427.

[1]The Penn Discourse Treebank Prasad et al. 2008 could also be considered as a corpus with partial dependency structures.

Proceedings of the SIGDIAL 2013 Conference, pages 2–11,
Metz, France, 22-24 August 2013. ©2013 Association for Computational Linguistics

function like logical operators on the meaning of their arguments. Hence their exact scope has great semantic impact on the phenomena we have mentioned, in exactly the way the relative scope of quantifiers make a great semantic difference in first order logic. By concentrating on exact meaning representations, however, the syntax-semantics interface becomes quite complex: as happens with quantifiers at the intra sentential level, discourse relations might semantically require a scope that is, at least a priori, not determined by syntactic considerations alone and violates surface order (see s_2).

Other theories like Polanyi's Linguistic Discourse Model (LDM) of Polanyi 1985; Polanyi and Scha 1984, and DLTAG Webber et al. 1999 explicitly adopt a syntactic point of view, and RST with strongly constrained (tree-shaped) structures is subject to parsing approaches duVerle and Prendinger 2009; Sagae 2009; Subba and Di Eugenio 2009 that adhere to the syntactic approach in adopting decoding strategies of syntactic parsing. In such theories, discourse structure representations, subject to syntactic constraints (*e.g.* dominance of spans of text one over another) respect surface order but do not always and unproblematically yield a semantic interpretation that fits intuitions. According to Marcu 1996, an RST tree is not by itself sufficient to generate desired predictions; he employs the *nuclearity principle*, NP, as an additional interpretation principle on scopes of relations.

We focus on two theories: RST, which offers the model for the annotations of the RST treebank Carlson, Marcu, and Okurowski 2002 and the Potsdam commentary corpus Stede 2004, and on SDRT, which counts several small corpora annotated with semantic scopes, Discor Baldridge, Asher, and Hunter 2007 and Annodis Afantenos et al. 2012. We describe these theories in section 2. We will also compare these two theories to dependency tree representations of discourse Muller et al. 2012. Section 3 introduces a language for describing semantics scopes of relations that is powerful enough to: i) compare the expressiveness (in terms of what different scopes can be expressed) of the different formalisms considered; ii) give a formal target language that will provide comparable interpretations of the different structures at stake. Section 4 discusses Marcu's nuclearity principle and proposes an alternative way to interpret an RST tree as a set of different possible scopes expressed in our language. Section 5 provides intertranslability results between the different formalisms. Section 6 defines a measure of similarity over discourse structures in different formalisms.

2 Discourse formalisms

These formalisms we introduce here all require the input text to be segmented into elementary units (EDUs). The definition of what an EDU is varies slightly with the formalism, but roughly corresponds to the clause level in RST, SDRT and other theories. We assume a segmentation common to the different formalisms and use examples with a non controversial and intuitive segmentation.

Rhetorical Structure Theory (RST), the theory underlying the RST-Treebank is the most used corpus for discourse parsing, cf. duVerle and Prendinger 2009, Subba and Di Eugenio 2009, *inter alia*.

In its Mann and Thompson 1987 formulation, RST builds a descriptive tree for the discourse by the recursive application of *schemata* in a bottom-up procedure. Each schema application ideally reflects the most plausible relation the writer intended between two contiguous spans of text, as well as hierarchical information about the arguments of the relation, distinguishing between *nuclei* as essential arguments of a relation and *satellites* as more contingent parts. The set of RS Trees is inductively defined as follows:

1- An EDU is a RS Tree.

2- if R is a nucleus-statellite relation symbol, s_1 and s_2 are both RS Trees with contiguous spans (the leftmost leaf in s_2 is textually located right after the rightmost one in s_1), and $\langle a_1, a_2\rangle \in \{\langle N, S\rangle; \langle S, N\rangle\}$ then $R(t_1\_a_1, t_2\_a_2)$ is an RS Tree.

3- if R is a multinuclear relation symbol and $\langle s_1, \ldots, s_n\rangle$ are n RS Trees with contiguous spans then $R(s_1\_N, \ldots, s_n\_N)$ is an RS Tree.

Following Mann and Thompson 1987 a complete RS tree makes explicit the content the author intended to communicate. RS Trees are graphically represented Marcu 1996 with intermediate nodes labelled with relation names, leaves with symbols referring to EDUs, and edges with nucleus/satellite distinctions.

Segmented Discourse Representation Theory (SDRT), our second case-study theory, inherits a framework from dynamic semantics and enriches it with rhetorical relations. The set of SDRSs is inductively defined as follows:

Assume a set of rhetorical relations $\mathcal{R}$, distinguished between coordinating and subordinating relations.

- Any EDU is an SDRS.

- Any Complex Discourse Unit (CDU) is a SDRS.

- a CDU is an acyclic labelled graph (A, E) where every node is a discourse unit (DU) or SDRS and each labelled edge is a discourse relation such that:

(i) every node is connected to some other node;

(ii) no two nodes are linked by subordinating and coordinating relations,

(iii) given EDUs $a_1, \ldots, a_{n+1}$ in their textual order that yield a CDU $(A, E) = G$, each EDU a_{j+1} $j < n$ is linked either: (a) to nodes on the right frontier of the CDU $G*$ a subgraph of G constructed from $a_1, \ldots, a_j$; or (b) to one or more nodes in $G' = (A', G')$, a subgraph of G, which linked to one or more nodes on the right frontier of the graph $G*$, and where G' is constructed from a subset of $a_{j+2}, \ldots a_n$.

The right frontier of a graph G consists of the nodes a that are not the left arguments to any coordinating relation and for which if any node b is linked to some node dominating a, then there is a path of subordinating

relations from b to a.

A Segmented Discourse Representation Structure (SDRS), is assigned a recursively computed meaning in terms of context-change potential (relation between pairs of ⟨ world, assignation function ⟩) in the tradition of dynamic semantics. The semantics of a complex constituent is compositionally defined from the semantics of rhetorical relations and the interpretation of its subconstituents. In the base case of an EDU, the semantics is given in dynamic semantics.

We also consider dependency trees (DTs). Muller et al. 2012 derive DTs from the SDRSs of the ANNODIS corpus to get a reduced search space, simplifying automated discourse parsing. A DT is an SDRS in which there are no CDUs and there is a unique arc between any two nodes. Muller et al. 2012 provide a procedure from SDRSs to DTs, which we slightly modify to respect the Frontier Contraint that they use. ζ works in a bottom-up fashion replacing every CDU X that is an argument of a rhetorical relation in γ by their top-most immediate sub-constituent which do not appear on the right of any relation in X, or distributing the top relation when necessary to preserve projectivity. To give a simple example: $\zeta(R([R'(a,[R''(b,c)])],d)) = \zeta(R([R'(a,b) \wedge R''(b,c)],d)) = R(a,d) \wedge R'(a,b) \wedge R''(b,c)$. (1) provides a more complicated example we discuss in Section 6).

3 Describing the scope of relations

We provide here a language expressive and general enough to express the structures of the 3 theories. All our case-study theories involve structures described by a list of rhetorical relations and their arguments. Two things may vary: first, the nature of the arguments. SDRT for instance, introduces *complex constituents* as arguments of relations (*e.g.* $\begin{cases} \pi : R_{subord}(b,c) \\ R_{subord}(a,\pi) \end{cases}$), which finds a counterpart within RS Trees, where a relation may directly appear as argument of another ($R(a_N, R(b_N, c_S)_S)$) but not within dependency trees. Second, the set of constraints that restrict the possible lists of such relations can vary across theories (e.g. right frontier, or requirement for a tree structure).

To deal with the first point above, we remark that it suffices to list, for each instance of a discourse relation, the set of *elementary* constituents that belong to its left and right scope in order to express the three kinds of structures. We do this in a way that an isomorphic structure can always be recovered. Models of our common language will be a list of relation instances and elementary constituents, together with a set of predicates stating what is in the scope of what. As for the second point, we axiomatize each constraint in our common language, thereby describing each of the 3 types of discourse structures as a theory in our language.

Our language contains only binary relations. Among discourse formalisms, only RST makes serious (and empirical) use of $n-$ary discourse relations. Nevertheless, such RST structures are expressible in our framework, if we assume certain semantic equivalences. RST allows for two cases of non-binary trees: (i) nucleus with n satellites, each one linked to the nucleus by some relation R_n. Such a structure is semantically equivalent to the conjunction of n-binary relations R_n between the nucleus and the nth satellite, which is expressible in our framework. (ii) RST also allows for n-ary multinuclear relations such as *List* and *Sequence*. In our understanding, multinuclear relations $R(a_1, \dots a_n)$, essentially serve a purpose of expressiveness, and such an n-ary tree is an equivalent to the split non-tree shaped structure $R(a_1, a_2) \wedge R(a_2, a_3) \dots R(a_{(n-1)}, a_n)$. This seems clear for the *Sequence* relation, which states that $a_1 \dots a_n$ are in temporal sequence and can be equivalently formulated as "each a_i precedes a_{i+1}". This might appear less obvious for the *List* relation. The semantics (as it appears on the RST website `http://www.sfu.ca/rst/`) of this relation requires the a_i to be "comparable", and as far as this is a transitive property, we can split the relation into a set of binary ones.

Formally, our scope language L_{scopes} is a fragment of that of monadic second order logic with two sorts of individuals: relation instances (i), and elementary constituants (l). Below, we assume $\mathcal{R}$ is the set of all relation names (elaboration, narration, justification, ...).

Definition 1 (Scoping language)**.** Let S be the set $\{i, l\}$. The set of primitive, disjoint types of L_{scopes} consists of i, l and t (type of formulae). For each of the types in S, we have a countable set of variable symbols V_i (V_l). Two additional countable sets of variable symbols $V_{\langle i,t\rangle}$ and $V_{\langle l,t\rangle}$ range over sets of individuals. These four sets of variable symbols are pairwise disjoint.

The alphabet of our language is constituted by V_i, V_s, a set of predicates, equality, connector and quantifier symbols. The set of predicate symbols is as follows:

1) For each relation symbol r in $\mathcal{R}$, L_R is a unary predicate of type $\langle i,t\rangle$—i.e., $L_R : \langle i,t\rangle$.

2) unary predicates, *sub*, *coord* and $sub^{-1} : \langle i,t\rangle$.

3) binary predicates $\in_l$ and $\in_r : \langle i,l,t\rangle$.

4) two equality relations, $=_s : \langle s,s,t\rangle$ for $s \in \{i,l\}$.

Logical connectors, and quantifiers are as usual. The sets of terms Γ_i, Γ_l and Γ_t are recursively defined: 1. $V_i \subseteq \Gamma_i$, $Var_l \subseteq \Gamma_l$. 2. For $v \in V_{s,t}$, $v : \langle s,t\rangle$. 3. For each symbol σ of type $\langle u_1, \dots, u_n\rangle$ in the alphabet, for all $(t_1, \dots, t_{n-1}) \in \Gamma_{u_1} \times \dots \times \Gamma_{un-1}$, $\sigma[t_1, \dots, t_{n-1}] \in \Gamma_{u_n}$. Γ_t is the set of well formed formulae of the scope language.

The predicates $\in_l$ and $\in_r$ take a relation instance r of type i and a elementary constituent x of type l as arguments. Intuitively, they mean that x has to be included in the left (for $\in_l$) or right (for $\in_r$) scope of r. For each relation symbol R such as *justification* or *elaboration*, the predicate L_R takes a relation instance r has argument and states that r is an instance of the rhetorical relation R. Predicates *sub*, *coord* and sub^{-1} apply to a relation instance r, respectively specifying that r's left argument hierarchically dominate its right argument, that

both are of equal hierarchical importance, or that the left one is subordinate to the right one.

Definition 2 (Scope structure and Interpretation)**.** A *scope structure* is an L_{scopes}-structure $\mathcal{M} = \langle D_i, D_l, |.|^{\mathcal{M}}\rangle$. D_i and D_l are disjoint sets of individuals for the sorts i and l respectively, and $|.|^{\mathcal{M}}$ assigns to each predicate symbol P of type $\langle u_1, \ldots, u_n, t\rangle$ a function $|.|^P : D_{u_1} \times \cdots \times D_{u_n} \mapsto \{0, 1\}$. Variables of type $\langle i, t\rangle$ are assigned subsets of D_i and similarly for variables of type $\langle l, t\rangle$, The predicates $=_i$ and $=_s$ are interpreted as equality over D_i and D_l respectively.

The interpretation $[\![\cdot]\!]_v^{\mathcal{M}}$ of a formula $\phi \in \Phi_S$ is the standard interpretation of a monadic second order formula w.r.t to a model and a valuation (interpretation of first order quantifiers and connectors is as usual, quantification over sets is over all sets of individuals). Validity $\models$ also follows the standard definition.

These scope structures offer a common framework for different discourse formalisms. Given one of the three formalisms, we say that two structures S_1 and S_2 are equivalent iff there is an encoding from one structure into a scoped structure or set of scoped structures and a decoding back from the scoped structure or set of scoped structures into S_2

Fact 1. One can define two algorithms I and E such that:

- from a given structure s which is a RS Tree, a SDRS or a DT, I computes a scope structure $I(s)$.
- given such a computed structure, E allow to retrieve the original structure s ($E(I(s)) = s$).

RST Encoding and Decoding To flesh out I and E for RST, we need to define dominance. Set $lArgs(r) = \{e \in D^l \mid (r, e) \in |\in_l|^{\mathcal{M}}\}$; $rArgs(r)$ is defined analogously (where $\in_r$ replaces $\in_l$). The left and right dominance relations $\sqsubseteq_l$ and $\sqsubseteq_r$ are defined as follows: $r \sqsubseteq_l r'$ iff $(Args(r) \subseteq lArgs(r'))$.
- $r \sqsubseteq_l r' \leftrightarrow \forall z : l((z \in_l r) \vee z \in_r r)) \rightarrow z \in_l r')$ with $r \sqsubseteq_r r'$ defined analogously.
Dominance $\sqsubseteq$ is: $\sqsubseteq = \sqsubseteq_l \cup \sqsubseteq_r$.
- $lArgs(r, X) \leftrightarrow \forall z : l(z \in_l r) \leftrightarrow z \in X)$, with $rArgs(r, X)$ similar and
-$Args(r, X) \leftrightarrow \forall z : l((z \in_l r) \vee z \in_r r)) \leftrightarrow z \in X)$.

The NS, NN and NS schemes of RST will be respectively encoded by the predicates *sub*, *coord* and sub^{-1}. We proceed recursively. If t is an EDU e, return $M_t = \langle D_i = \emptyset, D_l = \{e\}, \epsilon\rangle$ where ϵ is the interpretation that assigns the empty set to each predicate symbol. If the root of t is a binary node instantiating a relation $R(t_{1_{a_1}}, t_{2_{a_2}})$, let $T_r \in \{sub, coord, sub^{-1}\}$ be the predicate that encodes the schema a_1a_2, let $M_{t_1} = \langle D_i^1, D_l^1, |.|^1\rangle$ and $M_{t_2} = \langle D_i^2, D_l^2, |.|^2\rangle$. The algorithm returns $M_t = \langle D_i^1 \cup D_i^2 \cup \{r\}, D_l^1 \cup D_l^2, |.|^{M_t}\rangle$ where r is a 'fresh' relation instance variable not in D_i^1 or D_i^2, and $|.|^{M_t}$ is updated in the appropriate fashion to reflect the left and right arguments of r. Finally, if the root of t is an n-ary node, split it into a sequence of binary relation $R_1(t_1, t_2), R_2(t_2, t_3), \ldots$, proceed to recursively compute the scope-structures M_i for each of the relations using 2 (take care to introduce a 'fresh' relation instance individual for each relation of the sequence), then return the union of the models M_i.

RST Decoding Given a **finite** scope structure $\mathcal{M} = \langle D^i, D^l, |.|^{\mathcal{M}}\rangle$, for each relation instance r compute the left arguments of r and its right arguments. We then identify $L(r)$, the unique relation symbol R such that $r \in |L_R|^{\mathcal{M}}$. If that fails, the algorithm fails. Similarly retrieve the right nuclearity schema from the adequate predicate that applies to r. Then compute the dominance relations for r. If the input structure $\mathcal{M} = I(t)$ for some RS Tree t then there is at least one maximal relation instance for the dominance relation. If t the root node of t is a binary relation, there is exactly one maximal element in the dominance relation. If there is none, then we return fail. If there is exactly one, recursively compute the two RS Trees obtained from the models computed from the left and right arguments and descendants of r. If there is more than one, the root node of the encoded RS Tree was a n-ary relation and one has to reconstruct the n-ary node if that is possible; if not the algorithm fails (but that means the input structure was not obtained from a valid RS Tree).

SDRT Encoding and Decoding: This is similar to the RST encoding and decoding; for the encoding algorithm, we proceed recursively top down. A SDRS s is a complex constituent that contents a graph $g = \langle V, E\rangle$ whose edges are relations holding between sub-constituents, simple or complex as well. First come up with an encoding of the set E of all edges that hold between two sub-constituents of s, *i.e.* a structure $\mathcal{M} = \langle D_i = E_i, D_l = V, \{L_R\}, \in_l, \in_r\rangle$, where, for each edge $e \in E_i$, L_R encodes its relation type, and $\in_l{}^1$ and $\in_r{}^1$ consists of all the pairs (x, e) of left and right nodes x of the edges $e \in E$. Finally, for each complex immediate sub-constituent of s in D_l, update $\mathcal{M}$ as follows: for c such a subconsituent, recursively compute its encoding M_c, then add everything of M_c to M, finally remove c from M but add instead for each relation r scoping over c to the right (left), all the pairs $\{(r, x) \mid x$ is a constituent in $M^c\}$. The decoding works again similarly to the one for RST, top-down once again: one recursively retrieves immediate content of the current complex constituent at each level then moves to inner constituents.

DT: Dependency trees are syntactically a special case of SDRSs; there is only one CDU whose domain is only EDUs.

The scope language allows us to axiomatize three classes of scope structures corresponding to RS Trees, SDRSs and DTs. Not every scope structure will yield a RS Tree when fed to the RST decoding algorithm, only those obtainable from encoding an RS tree. As not all scope structures obey these axioms, our language is

strictly more expressive than any of these discourse formalisms.

As an example of an axiom, the following formula expresses that a relation cannot have both left and right scope over the same elementary constituent:

Strong Irreflexivity:

$$\forall r : i \forall x : l \neg(x \in_l r \ \wedge \ x \in_r r)) \qquad (A_0)$$

Strong irreflexivity entails irreflexivity; a given relation instance cannot have the same (complete) left and right scopes. All discourse theories validate A_0.

In the Appendix, we define left and right strong dominance relations $\sqsubseteq_{l(r)}$ as well as n-ary RS trees and CDUs of SDRT. We exploit these facts in the Appendix to express axioms (A1-A9) that axiomatize the structures corresponding to RST, SDRT and DTs. Axiom A_1 says that every discourse unit is linked via some discourse relation instance. Axiom A_2 insures that all our relation instances have the right number of arguments; Axioms A_3 and A_4 ensure acyclicity and no crossing dependencies. A_5a and A_5b restrict structures to a tree-like dominance relation with a maximal dominating element, while A_6 defines the Right Frontier constraint for SDRT, and A_7 fixes the domain for SDRT constraints on CDUs. A_8 ensures that no coordinating and subordinating relations have the same left and right arguments, while A_9 provide the restrictions needed to define the set of DTs. We use the encoding and decoding maps to show:

Fact 2.

1. The theory T_{RST}=$\{A_0, A_1, A_2, A_3, A_4, A_5a, A_5b, A_8\}$ characterizes RST structures in the sense that:

   - E applied to any structure M such that $M \models T_{RST}$ yield an RST Tree.

   - for any RST Tree t, $I(t) \models T_{RST}$.

2. The theory T_{SDRT}=$\{A_0, A_1, A_2, A_3, A_6, A_7, A_8\}$ similarly characterizes SDRSs.

3. The theory T_{DT}=$T_{SDRT} \cup \{A_9a, A_9b\}$ similarly characterizes Dependency Trees structures.

4 Different Interpretations of Scope

The previous section defined the set of scope structures as well as the means to import, and then retrieve, RS trees, DTs, or SDRs into, and from, this set. Some of these scope structures export both into RST and SDRT, yielding a $1-1$ correspondence between a subset of SDRT and RST structures. But what does this correspondence actually tell us about these two structures? In mathematics, the existence of an isomorphism relies on a bijection that *preserves* structure. Our correspondence preserves the *immediate interpretation* of the semantic scopes of relations.

Immediate Interpretation Consider a scope structure $\mathcal{M}$ (validating A_0, A_1, A_2). The predicates $lArgs(r)$ and $rArgs(r)$ are the sets of all units in the left or right scope of a relation instance r. Whether r, labelled by relation name R holds of two discourse units or not in $\mathcal{M}$, depends on the semantic content of its left and right arguments, recursively described by $lArgs(r)$ and all relations r' such that $r' \sqsubset_l r$, and $rArgs(r)$ and all relations r' such that $r' \sqsubset_r r$. Algorithm I computes what we call the *immediate* interpretation of an input structure. Intuitively, in this interpretation the semantic scope of relations is directly read from the structures themselves; a node $R(t_1, t_2)$ in a RS Tree expresses that R holds between contents expressed by the whole substructures t_1 and t_2. Similarly, for SDRT and DTs, immediate interpretation of an edge $\pi_1 \rightarrow_R \pi_2$ is that R holds between the whole content of π_1 and π_2.

While this immediate interpretation is standard in SDRT, it is not in RST. Consider again (1) from the introduction or:

(2) [In 1988, Kidder eked out a $ 46 million profit,]$_{31}$ [mainly because of severe cost cutting.]$_{32}$ [Its 1,400-member brokerage operation reported an estimated $ 5 million loss last year,]$_{33}$ [although Kidder expects to turn a profit this year]$_{34}$ (RST Treebank, wsj_0604).

(3) [Suzanne Sequin passed away Saturday at the communal hospital of Bar-le-Duc,]$_3$ [where she had been admitted a month ago.]$_4$ [...] [Her funeral will be held today at 10h30 at the church of Saint-Etienne of Bar-le-Duc.]$_5$ (annodis corpus).

These examples involve what are called *long distance attachments*. (2) involves a relation of contrast, or comparison between 31 and 33, but which does not involve the contribution of 32 (the costs cutting of 1988). (3) displays something comparable. A causal relation like result, or at least a temporal narration holds between 3 and 5, but it should not scope over 4 if one does not wish to make Sequin's admission to the hospital a month ago a consequence of her death last Saturday. Finally in (1) C_4 elaborates on C_1, but not on the fact that C_1 is attributed to chief Garcia, so the corresponding elaboration relation should not scope over C_3.

It is impossible however, to account for long distance attachment using the immediate interpretation of RST trees. (2), for instance, also involves an explanation relation between 31 and 32, which should include none of 33 or 34 in its scope. Since 31 is in the scope of both the explanation and the contrast relation, Axiom A_5a of the previous section entails than an RST tree involving the two relations has to make one of the two relations dominates the other.

Marcu's Nuclearity Principle (NP) Marcu 1996 provides an alternative to the immediate interpretation and captures some long distance attachments Danlos 2008; Egg and Redeker 2010. According to the NP, a rela-

tion between two spans of text, expressed at a node of a RS Tree should hold between the most salient parts of these spans. *Most salient part* is recursively defined: the most salient part of an elementary constituent is itself, for a multinuclear relation $R(t_{1N}, \ldots, t_{kN})$ its most salient part is the union of the most salient parts of the t_i[2]. Following Egg and Redeker 2010, the NP, or *weak* NP is a constraint on which RST trees may correctly characterize an input text; it is not a mechanism for computing scopes. Given their analysis of (1) given in the introduction, NP entails that $Elab_1$ holds between C_1 and C_4, accounting for the long distance attachment, and that Attribution holds between C_1 and C_4 which meets intuition in this case. There is however no requirement that *Attribution* do *not* hold between the wider span $[C_1, C_2]$ and C_3, as there is no requirement that $Elab_1$ does not hold between $[C_1, C_2, C_3]$ and C_4. In order to accurately account for (1), the former must be true and the latter false.

However, this interpretation of NP together with an RST tree does not determine the semantic scope of all relations. Danlos 2008 reformulates NP as a *Mixed* Nuclearity Principle (MNP) that outputs determinate scopes for a given structure. The MNP requires for a given node, that the most salient parts of his daughters furnish the **exact** semantic scope for the relation at that node. The MNP transforms an RST tree t into a scope structure $\mathcal{M}_t$, which validates $A_0 - A_3$ but also A_6.[3], A_7 and A_8. Hence $\mathcal{M}$ could be exported back to SDRT and the MNP would yield a translation from RST-trees to SDRSs.

But when applied to the RST Treebank, the MNP yields wrong, or at least incomplete, semantic scopes for intuitively correct RS Trees. The mixed principle applied to the tree of s_1 gives the Attribution scope over C_1 only, but not C_2, which is incorrect. Focusing on the attribution relation which is the second most frequent in the RST Treebank, we find out that, regardless of whether we assign Attribution's arguments S and N or N and S, this principle makes wrong predictions 86% of the time in a random sampling of 50 cases in which we have attributions with multi-clause second argument spans. Consider the following example from the RST Treebank:

(4) [Interprovincial Pipe Line Co. said]$_1$ [it will delay a proposed two-step, 830 million Canadian-dollar [(US$705.6 million)]$_3$ expansion of its system]$_2$ [because Canada's output of crude oil is shrinking.]$_4$

Applied to the annotated RS Tree for this example (figure 1), the MNP yields an incorrect scope of the attribution relation over 2 only, regardless of whether the attribution is annotated *N-S* or *S-N*. The idea behind the weak NP provides a better fit with intuitions. The principle gives *minimal* semantic requirements for scoping relations; everything beyond those requirements is left *underspecified*. We formalize this as the *relaxed* Nuclearity Principle (RNP), which does not compute one structure where each relation is given its exact scope, but **a set** of such structures.

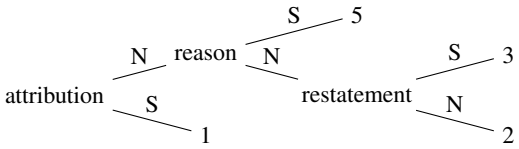


Figure 1: Annotated RST Tree for example (4).

The target structures are not trees any more, but we want them to still reflect the dominance information present in the RS Tree. We therefore define a notion of *weak dominance* over structures of the scoping language: for two sets of constituents, $X \preceq Y$ iff $X \subseteq Y$ or there is a subordinating relation whose left argument is X and right one Y. Weak dominance is given by transitive closure $\preceq^*$ of $\preceq$. For two relations, $r \preceq^*_l r'$ iff the left argument of r weakly dominates both arguments of r'. $\preceq^*_r$ is symmetrically defined. Finally, structures computed by the RNP have to validate the weakened version of A_5: if two relations scope over the same elementary constituent one has to weakly dominates the other. Let A_5^W denote this axiom.

Definition 3 (Relaxed Nuclearity Principle)**.** One can assign to an RS Tree t a formula of the scoping language $\phi_t = \exists\bar{x}\exists\bar{r}\psi_t \cup \Gamma_t$ such that:

1– ψ_t is a formula specifying that all individuals quantified in $\bar{x}$ and $\bar{r}$ are pairwise distinct, and that there is no other individuals that the ones just mentioned. ψ_t also specifies for each intermediate node n that the corresponding relation instance r_n is labelled with the adequate relation symbol R and relation type (subordinating if *N-S* ...).

2– Γ_t encodes the nuclearity principle applied to t: for all intermediate nodes n_i and n_j in t such that n_l is the left (resp. right) daughter of n_i, Γ_t specifies that n_i must scope to the left (resp. right) over the nucleus of n_j.

The interpretation $[\![t]\!]$ is defined as the set of structures $\mathcal{M}$ that validate ϕ_t and $A_0, A_1, A_2, A_3, A_5^W$ (they all have $|t|$ individuals, as fixed by ψ_t). Moreover, it can be shown that each model of this set validates T_{SDRT}; so we have a interpretation of an RS-Tree into a set of SDRSs.

5 Intertranslability between RST/DTs

DTs are a restriction of SDRSs to structures without complex constituents. So the ζ function of section 2

[2]Except for Sequence which only retains the most salient part of t_k

[3]That A_6 is valid in the resulting model is not immediate. Assume a multinuclear (coordinating) relation instance r has scope over x_n and x_{n+k} later in the textual order. Then it is impossible to attach with r' a later found constituent x_{n+k+l} to x_n alone, for it would require that x_{n+1} escapes the scope of r' from the MNP which it will not do by multinuclearity of r.

can transform distinct SDRSs transform into the same DT with a consequent loss of information.

$$\begin{array}{l} a \rightarrow_{R_1} \pi \\ \pi : b \rightarrow_{R_2} c \end{array} \quad | \; a \rightarrow_{R_1} b \rightarrow_{R_2} c \; | \quad \begin{array}{l} \pi \rightarrow_{R_2} b \\ \pi : a \rightarrow_{R_1} b \end{array} \qquad (1)$$

Each of the SDRSs above yields the same DT after simplification, namely the second one $a \rightarrow_{R_1} b \rightarrow_{R_2} c$.

The natural interpretation of a DT g describes the set of fully scoped SDRS structures that are compatible with these minimal requirements, *i.e* that would yield g by simplification. To get this set, every edge $r(x, y)$ in g, r, must be assigned left scope among the *descendants* of x in g (and right scope among those of y); this is a consequence of i) x and y being *heads* of the left and right arguments of r and ii) the SDRSs that are compatible with g do not admit relations with a right argument in one constituent and a left one outside of it.

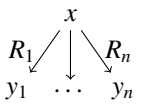
Definition 4. Assume that we map each node[4] x of g into a unique variable $v_x \in V_l$ and each edge e into a unique variable symbol $r_e \in V_i$. Define $\bar{x}$ and $\bar{r}$ in an analogous way as in definition 3.

For a given dependency tree g, we compute a formula $\phi_g = \exists\bar{x}\exists\bar{r} \;\; \psi_g \cup \Gamma_g$ such that

- ψ_g is defined analogously as in definition 3, defining the set of relation instances and EDUs.
- Γ_g is the formula stating the minimal scopes for each relation instance: for all edge in $e = R(x, y)$ in g, Γ_g entails i) r_e has v_x in its left scope and v_y in its right scope and ii) let $Des(x)$ be the set of variable symbols for all the descendants of x in g, Γ_g entails that if r_e has left scope over some v_z then v_z is in $Des(x)$ (symmetrically for y and right scope).

The interpretation $[\![g]\!]$ of a DT is: $\{\mathcal{M} \mid \mathcal{M} \models \phi_g, A_0\text{-}A_3, A_6, A_7\}$. The DT $a \rightarrow_{R_1} b \rightarrow_{R_2} c$ for instance, is interpreted as a set of three structures isomorphic to the ones in (1) above.

We now relate DTs to RS Trees interpreted with the RNP. To this aim, we focus on a restricted class of DTs, those who involve i) coordinating chains of 3 edus or more only if they involve a single coordinating relation: $x_1 \rightarrow_{R_1} x_2 \rightarrow_{R_2} \cdots \rightarrow_{R_{n-1}} x_n$ may appear only for $n > 2$ if all the R_i are the same coordinating relation, and ii) subordinating nests of 3 edus or more only if they involve a single subordinating relation:



is allowed for $n > 1$ only if all R_i are labelled with the same subordinating relation.

This restricted class of DTs corresponds exactly with the set of RS-Trees interpreted with the RNP, provided that we restrict the interpretation of a DT in the following way: a principle called *Continuing Discourse Pattern*, CDP Asher and Lascarides 2003 must apply, who states that whenever a sequence of coordinating relation R^i_c originates as a node which appear to be also in the right scope of a subordinating relation R_s, R_s must totally include all the R^i_c in its right scope. A second principle is required, who states that whenever two subordinating relations $R0_s$ and R'_s originate at the same node in the DT, and the right argument of R'_s is located after the right argument of R_s, any structure in the interpretation of the DT must verify $R'_s \preceq_l R_s$. The translation needs these requirements to work, because: i) with the NP a relation scoping over a multinuclear one must includes all the nucleus in RST, and ii)a node in a RS Tree cannot scope over something that is not its descendant). Let CDP$^+$ denote these requirements.

Using the restricted interpretation of a DT g; $[\![g]\!]^{CDP} = \{\mathcal{M} \mid \mathcal{M} \models A_0\text{-}A_3, A_6, A_7, \text{CDP}^+\}$, we transform an RS Tree t into a dependency graph $\mathcal{G}(t)$ such that $[\![t]\!] = [\![\mathcal{G}(t)]\!]^{CDP}$:

Definition 5 (RS Trees to dependency graphs)**.** The translation $\mathcal{G}$ takes a RS Tree t as input and outputs a pair $\langle G, n\rangle$, where $G = \langle Nodes, Edges\rangle$ is the corresponding dependency graph, and n an attachment point used along the recursive definition of $\mathcal{G}$.

- If t is an EDU x then $(G)(t) = \langle(\{x\}, \{\}), x\rangle$.
- If $t = R(t_{1N}, t_{2S})$ then let $\langle G_1, n_1\rangle = \mathcal{G}(t_1)$ and $\langle G_2, n_2\rangle = \mathcal{G}(t_2)$.

  $$\mathcal{G}(t) = \langle(G_1 \cup G_2 \cup \{R_{subord}(n_1, n_2))\}; n_1\rangle$$

- If $t = R(t_{1S}, t_{2N})$ then $\mathcal{G}(t) = \mathcal{G}(R(t_{2N}, t_{1S}))$
- If $t = R(t_{1N}, \ldots, t_{kN})$ (multinuclear), let $\langle G_i, n_i\rangle = \mathcal{G}(t_i)$, let G be the result of adding a chain $n_1 \rightarrow_{R_{coord}} \cdots \rightarrow_{R_{coord}} n_k$ to the union of the G_i,

  $$\mathcal{G}(t) = \langle G; n_1\rangle$$

- If t is a nuclear satellite relation with several satellites $R(t_{1S}, \ldots t_{j_N}, \ldots t_{kS})$, compute the G_i has in the previous case, then add to the union of the G_i the nest of $k-1$ subordinating relations R linking n_j to each of the n_i, $i \neq j$.

Recall RS Tree (s_1). Applying $\mathcal{G}$ to this tree yields the dependency tree (s_3): $Elab_1(C_1, C_2) \wedge \text{Attr}(C_1, C_3) \wedge Elab_2(C_1, C_4)$. $[\![s_3]\!]$ supports any reading of (s_1) provided by RNP, but also an additional one where *Attr* scopes over $[C_1, C_2, C_4]$. This is however forbidden by CDP+ for C_4 is after C_3 in the textual order but $Elab(C_1, C_4) \not\preceq_l Attr(C_1, C_3)$.

6 Similarities and distances

The framework we have presented yields a notion of similarity that applies to structures of different formalisms. To motivate our idea, recall example (1); the structure in (s_3) in which Attribution just scopes over C_1 differs from the intuitively correct interpretation only in that Attribution should also scope over C_2

[4] Recall that unlike RS Trees, DTs have EDUs as nodes and relations as edges.

as in (s_2), while a structure that does this but in which C_3 is in the scope of the Elaboration relation is intuitively further away from the correct interpretation.

Our similarity measure *Sim* over structures $\mathcal{M}_1$ and $\mathcal{M}_2$ assumes a common set of elementary constituents and a correspondence between relation types in the structures. We measure similarity in terms of the scopes given to the relations. The intuition, is that given a map f from elements of relation instances in $\mathcal{M}_1$ relation instances in $\mathcal{M}_2$, we achieve a similarity score by counting for each relation instance r the number of EDUs that are both in the left scope of one element of r and in $f(r)$, then divide this number by the total number of diffrents constituents in the left scope of r_1 and r_2, and do the same for right scopes as well. The global similarity is given by the correspondence which yields the best score.

Given a relation $r_1 \in \mathcal{M}_1$ and a relation $r_2 \in \mathcal{M}_2$, let $\delta(r_1, r_2) = \begin{cases} 1 \text{ if } r_1 \text{ and } r_2 \text{ have the same label} \\ 0 \text{ otherwise} \end{cases}$. Define $C_l(r_1, r_2) = |\{x : l \mid \mathcal{M}_1 \models x \in_l r_1 \wedge \mathcal{M}_2 \models x \in_l r_2\}|$, the number of constituents over which r_1 and r_2 scope and $D_l(r_1, r_2) = |\{x : l \mid \mathcal{M}_1 \models x \in_l r_1 \vee \mathcal{M}_2 \models x \in_l r_2\}|$. Define C_r and D_r analogously and assume that $\mathcal{M}_1$ has less relation instances than $\mathcal{M}_2$. Let $Inj(D_i^1, D_i^2)$ be the set of injections of relations instances of $\mathcal{M}_1$ to those of $\mathcal{M}_2$.

$$Sim(\mathcal{M}_1, \mathcal{M}_2) = \frac{1}{2Max(|\mathcal{M}_1|, |\mathcal{M}_2|)} \times \underset{f \in Inj(D_i^1, D_i^2)}{Max} \sum_{r:i} \delta(r, f(r)) \times \left(\frac{C_l(r, f(r))}{D_l(r, f(r))} + \frac{C_r(r, f(r))}{D_r(r, f(r))}\right)$$

If M_2 has more relation instances, Invert arguments and use the definition above. If they have same number of instances, both directions coincide.

$$d(\mathcal{M}_1, \mathcal{M}_2) = 1 - Sim(\mathcal{M}_1, \mathcal{M}_2)$$

For a discourse structure $\mathcal{M}$, $Sim(\mathcal{M}, \mathcal{M}) = 1$; Sim ranges between 1 and 0. d is a Jaccard-like metric obeying symmetry, $d(x, x) = 0$ $d(x, y) \neq 0$ for $x \neq y$, and the triangle equality. One can further define the maximal or average similarity between any pair of structures of two sets S_1 and S_2. This gives an idea of the similarity between two underspecified interpretations, such as the ones provided by RNP of section 4. For example, the maximal similarity between (s_2) interpreted as itself (immediate interpretation) and a possible scope structure for the DT (s_3), interpreted with the underspecified $[\![\,]\!]$ of section 5, is 7/12. It is provided by the interpretation of (s_3) where Attr is given left scope over C_1, C_2, C_4, $Elab_1$ holds between C_1 and C_2, and the second *Elab* fails to match the continuation of (s_3). $sim([\![s_2]\!], [\![\zeta(s_2)]\!]) = 7/12$ also, because ζ must distribute $[2, 4]$ in s_2 to avoid crossing dependencies; so $[\![\zeta(s_2)]\!] \cong [\![s_3]\!]$. The maximal similarity between the RS tree in (s_1) with RNP (or equivalently, (3) with $[\![\,]\!]^{CDP+}$) and (s_2) is 19/36, achieved when both C_1 and C_2 are left argument of Attr (though not C_4). With MNP, the similarity is 17/36.

Given our results in sections 4 and 5, we have:

Fact 3. (i) For any DT g without a > 3 length flat sequence and interpreted using CDP+, there an RS tree t interpreted with RNP such that $Sim(g, t) = 1$. (ii) For any RS tree with RNP there is a DT g such that $Sim(t, g) = 1$.

To prove (i) construct a model using Definition 4 and then use RST decoding. To prove (ii) construct a model given Definition 3 and use DT encoding. Our similarity measure provides general results for SDRSs and DTTs (and *a fortiori* SDRSs and RS trees) (See Appendix).

7 Related Work

Our work shares a motivation with Blackburn, Gardent, and Meyer-Viol 1993: Blackburn, Gardent, and Meyer-Viol 1993 provides a modal logic framework for formalizing syntactic structures; we have used MSO and our scope language to formalize discourse structures. While many concepts of discourse structure admit of a modal formalization, the fact that discourse relations can have scope over multiple elementary nodes either in their first or second argument makes an MSO treatment more natural. Danlos 2008 compares RST, SDRT and Directed Acyclic Graphs (DAGs) in terms of their *strong generative capacity* in a study of structures and examples involving 3 EDUS. We do not consider generative capacity, but we have given a generic and general axiomatization of RST, SDRT and DT in a formal interpreted language. We can translate any structure of these theories into this language, independent of their linguistic realization. We agree with Danlos that the NP does not yield an accurate semantic representation of some discourses. We agree with Egg and Redeker 2010 that the NP is rather a constraint on structures, and we formalize this with the relaxed principle and show how it furnishes a translation from RS trees to sets of scoped structures. Danlos's interesting correspondence between restricted sets of RST trees, SDRSs and DAGs assumes an already fixed scope-interpretation for each kind of structure: SDRSs and DAGs are naturally interpreted as themselves, and RS Trees are interpreted with the mixed NP Our formalism allows us **both** to describe the structures themselves and various ways of computing alternate scopes for relations.

With regard to the discussion in Egg and Redeker 2008; Wolf and Gibson 2005 of tree vs. graph structures, we show exactly how tree based structures like RST with or without the NP compare to graph based formalisms like SDRT. We have not investigated Graphbank here, but the scope language can axiomatize Graphbank (with A_0-A_3, A_8).

8 Conclusions

We have investigated how to determine the semantic scopes of discourse relations in various formalisms by

developing a canonical formalism that encodes scopes of relations regardless of particular assumptions about discourse structure. This provides a *lingua franca* for comparing discourse formalisms and a way to measure similarity between structures, which can help to compare different annotations of a same text.

References

Afantenos, S. et al. (2012). "An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus". In: *Proceedings of LREC 2012*. ELRA.

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

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Studies in Linguistics and Philosophy 50. Dordrecht: Kluwer.

Baldridge, J., N. Asher, and J. Hunter (2007). "Annotation for and Robust Parsing of Discourse Structure on Unrestricted Texts". In: *Zeitschrift fr Sprachwissenschaft* 26, pp. 213–239.

Bateman, J. and K. J. Rondhuis (1997). "Coherence relations : Towards a general specification". In: *Discourse Processes* 24.1, pp. 3–49.

Blackburn, P., C. Gardent, and W. Meyer-Viol (1993). "Talking about Trees". In: *EACL 6*, pp. 21–29.

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

Danlos, L. (2008). "Strong generative capacity of RST, SDRT and discourse dependency DAGSs". English. In: *Constraints in Discourse*. Ed. by A. Benz and P. Kuhnlein. Benjamins, pp. 69–95.

duVerle, D. and H. Prendinger (2009). "A Novel Discourse Parser Based on Support Vector Machine Classification". In: *Proceedings of ACL-IJCNLP 2009*. ACL, pp. 665–673.

Egg, M. and G. Redeker (2008). "Underspecified discourse representation". In: *PRAGMATICS AND BEYOND NEW SERIES* 172, p. 117.

– (2010). "How Complex is Discourse Structure?" In: *Proceedings of LREC'10*. Ed. by N. Calzolari et al. ELRA.

Hitzeman, J., M. Moens, and C. Grover (1995). "Algorithms for Analyzing the Temporal Structure of Discourse". In: *Proceedings of the 7th Meeting of the European Chapter of the Association for Computational Linguistics*, pp. 253–260.

Hobbs, J. R., M. Stickel, and P. Martin (1993). "Interpretation as Abduction". In: *Artificial Intelligence* 63, pp. 69–142.

Kehler, A. (2002). *Coherence, Reference and the Theory of Grammar*. CSLI Publications.

Lascarides, A. and N. Asher (1993). "Temporal Interpretation, Discourse Relations and Commonsense Entailment". In: *Linguistics and Philosophy* 16, pp. 437–493.

Mann, W. C. and S. A. Thompson (1987). "Rhetorical Structure Theory: A Framework for the Analysis of Texts". In: *International Pragmatics Association Papers in Pragmatics* 1, pp. 79–105.

Marcu, D. (1996). "Building up rhetorical structure trees". In: *Proceedings of the thirteenth national conference on Artificial intelligence - Volume 2*. AAAI'96. Portland, Oregon: AAAI Press, pp. 1069–1074. ISBN: 0-262-51091-X.

Muller, P. et al. (2012). "Constrained decoding for text-level discourse parsing". Anglais. In: *COLING - 24th International Conference on Computational Linguistics*. Mumbai, Inde.

Polanyi, L. (1985). "A Theory of Discourse Structure and Discourse Coherence". In: *Papers from the General Session at the 21st Regional Meeting of the Chicago Linguistics Society*. Ed. by P. D. K. W. H. Eilfort and K. L. Peterson.

Polanyi, L. and R. Scha (1984). "A Syntactic Approach to Discourse Semantics". In: *Proceedings of the 10th International Conference on Computational Linguistics (COLING84)*. Stanford, pp. 413–419.

Polanyi, L. et al. (2004). "A Rule Based Approach to Discourse Parsing". In: *Proceedings of the 5th SIGDIAL Workshop in Discourse and Dialogue*, pp. 108–117.

Prasad, R. et al. (2008). "The penn discourse treebank 2.0". In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, p. 2961.

Sagae, K. (2009). "Analysis of Discourse Structure with Syntactic Dependencies and Data-Driven Shift-Reduce Parsing". In: *Proceedings of IWPT'09*. ACL, pp. 81–84.

Sanders, T., W. Spooren, and L. Noordman (1992). "Toward a taxonomy of coherence relations". In: *Discourse processes* 15.1, pp. 1–35.

Stede, M. (2004). "The Potsdam Commentary Corpus". In: *ACL 2004 Workshop on Discourse Annotation*. Ed. by B. Webber and D. K. Byron. Barcelona, Spain: Association for Computational Linguistics, pp. 96–102.

Subba, R. and B. Di Eugenio (2009). "An effective Discourse Parser that uses Rich Linguistic Information". In: *Proceedings of HLT-NAACL*. ACL, pp. 566–574.

Webber, B. et al. (1999). "Discourse Relations: A Structural and Presuppositional Account Using Lexicalised TAG". In: *Proceedings of the 37th ACL Conference*. College Park, Maryland, USA: Association for Computational Linguistics, pp. 41–48. DOI: 10.3115/1034678.1034695.

Wolf, F. and E. Gibson (2005). "Representing Discourse Coherence: A Corpus Based Study". In: *Computational Linguistics* 31.2, pp. 249–287.

Appendix

In what follows, let $\sqsubset$ denotes the irreflexive part of $\sqsubseteq$ We assume that we have access to the textual order

of EDUs as a function f : EDUs $\rightarrow N$ with an associated strict linear ordering $<$ over EDUs. We also appeal to the notion of a chain over EDUs $\{x_1, x_2, \ldots x_n\}$ with a set of relation instances $r_1, \ldots, r_n\}$ *all of which are instances of an n-ary relation type*, of the form $x_1 \rightarrow^{r_1} x_2 \rightarrow^{r_2} \ldots \rightarrow^{r_n} x_n$ which can be defined in MSO. To handle RST relations with multiple satellites, we define a *nest*: $Nest(X, R)$ iff all $r \in R$ have the same left argument in X but take different right arguments in X. Finally, we define CDUs:

$$cdu(X, R) \leftrightarrow \exists r Args(r, X) \wedge \forall r' \, (\forall x \, x \in_r r' \rightarrow x \in X) \rightarrow r' \in R$$

Axiomatization

$$\forall x : l \, \exists r : i \quad (x \in_l r) \vee (x \in_r r) \qquad (A_1\text{:Weak Connectedness})$$

$$\forall r \exists x, y (x \in_r r) \wedge y \in_l r)) \qquad (A_2 \text{ :Properness of the relation})$$

$$\forall X : (l, t)(X \neq 0 \rightarrow \exists y \in X \, \forall n \neg y \in_l n \qquad (A_3 \text{ :Acyclicity or Well Foundedness})$$

No crossing dependencies using the textual order $<$ of EDUs:

$$\forall x, y, z, w((x < y < z < w) \rightarrow \forall m, n \neg (x \in_l n \wedge z \in_r n \wedge y \in_l m \wedge w \in_r m)) \qquad (A_4)$$

Tree Structures. Define $scopes(r, x) := x \in_l r \vee x \in_r r$.

$$\forall r, r' \, ((\neg(\exists X, R \, r, r' \in R \wedge \text{chain}(X, R) \wedge \text{nest}(X, R)) \wedge (\exists x \, scopes(r, x) \wedge scopes(r', x))) \rightarrow (r \sqsubseteq r' \vee r' \sqsubseteq r)) \qquad (A_5 a)$$

$$\forall R : (i, t) \exists ! r : i \, \forall r' \in R \;\; r' \sqsubseteq r \qquad (A_5 b)$$

Right Frontier:

$$\forall n, x_n, x_{n+1} \forall r \, ((x_{n+1} \in_r r) \rightarrow (x_n \in_l r) \vee (\neg x_n \in_l r \rightarrow \exists X, R(\text{chain}(X, R) \wedge \forall r'(r' \in R \rightarrow sub(r')) \wedge \exists y \in X \exists z \exists k \, \exists m, j \in R \, (scopes(j, y) \wedge acc(z, y) \wedge scopes(m, x_n) \wedge z \in_l k \wedge k \prec *x_{n+1})))) \qquad (A_6)$$

(The definition of SDRS accessibility *acc* is easy)

CDUs or EDUs and no overlapping CDUs:

$$\exists ! x : l \vee \exists X, R \, cdu(X, R) \wedge \forall X, Y, R, R' \;\; (cdu(X, R) \wedge cdu(Y, R') \rightarrow (R \cap R' \neq 0 \rightarrow (R \subseteq R' \vee R' \subseteq R)) \qquad (A_7)$$

The same arguments cannot be linked by subordinating and coordinating relations. The formal axiom is evident.

Finally, two axioms for restricting SDRSs to dependency trees:

$$\forall r \forall x, y((x \in_l r) \wedge y \in_l r)) \vee (x \in_r r) \wedge y \in_r r))) \rightarrow x = y \qquad (A_9 a : NoCDUs.)$$

$$\forall r \forall r' \forall X, Y (lArgs(r, X) \wedge rArgs(r, Y) \wedge lArgs(r', X) \wedge rArgs(r', Y)) \rightarrow r = r' \qquad (A_9 b \text{ :unique arc})$$

We note that as a consequence of $A_5 a$ and $A_5 b$ we have no danglers or contiguous spans:

$$\forall x, y, n \, (x \in_l n \wedge y \in_l n \wedge x \neq y) \rightarrow \neg \exists m \exists z \, (x \in_l m \wedge z \in_r m \wedge \neg(z \in_l n \vee z \in_r n))$$

We also note that $A_5 a$ and $A_5 b$ entail A_7, A_8 and $A_9 b$, though not vice-versa.

Fact 4. Where γ is any SDRS and $\zeta : SDRS \rightarrow DT$ as in section 2, set $R_1 = \{r : i \; : \; |\{x : M_\gamma \models x \in_l r)\}| > 1\}$, $R_2 = \{r : i \; : \; |\{x : M_\gamma \models x \in_r r)\}| > 1\}$, and $R_{\{x,y\}} = \{r | \exists r' : i (x \in_l r' \wedge y \in_r r' \wedge r' \neq r\}$. Assume the **immediate interpretation** of γ and $\zeta(\gamma)$:

$$Sim(\gamma, \zeta(\gamma)) = \frac{2|I| - |(R_1 \cup R_2) \cup \bigcup_{x,y \in D_l^2} X_{\{x,y\}}|}{2|I|} + \frac{1}{2|I|} \{\Sigma_{r \in R_1} \frac{1}{|x : M_\gamma \models x \in_l r)|} + \Sigma_{r \in R_2} \frac{1}{|x : M_\gamma \models x \in_r r)|}\}$$

Explanation: We suppose that I is the number of relation instances in the SDRS. ζ removes CDUs in an SDRS and attaches all incoming arcs to the CDUs to the head of the CDU. It also removes multiple arcs into any given node. So for any node m such that $|\{r : m \in_r r\}| = a > 1$, then the information contained in the $a - 1$ arcs will be lost. In addition ζ will restrict that one incoming arc that in the SDRS has in its scope all the elements in the CDU to just the head. So the scope information concerning all the other elements in the CDU will be lost.