

Towards a Discourse Relation Algebra for Comparing Discourse Structures

Charlotte Roze

Université Paris Diderot, Sorbonne Paris Cité, ALPAGE, UMR-I 001 INRIA,
charlotte.roze@linguist.jussieu.fr,
WWW home page: www.linguist.univ-paris-diderot.fr/~croze/

Abstract. We propose a methodology for building discourse relations inference rules, to be integrated into an algebra of these relations. The construction of these rules has as main objective to allow for the calculation of the discourse closure of a structure, i.e. deduce all the discourse relations it implicitly contains. Calculating the closure of discourse structures improves their comparison, in particular within the evaluation of discourse parsing systems. We present and illustrate the adopted methodology, taking as theoretical background the *Segmented Discourse Representation Theory* or SDRT (Asher and Lascarides, 2003).

1 Introduction

Discourse parsing, the task of producing a discourse structure for a given text, raises two crucial issues: building gold-standard corpora and evaluating discourse annotations. Building a gold-standard generally implies the merging of different annotations of the same text(s) and therefore the comparison of discourse structures produced by different annotators. Evaluating discourse annotations also requires to compare discourse structures generated by a system and those of the reference. But comparing discourse structures is not a trivial task: two discourse annotations of the same text can be different without any of them being wrong or incomplete. For example, the discourse in (1), which contains three elementary discourse units (π_1) , (π_2) and (π_3) , can receive two different annotations: an annotation A_1 , containing *Result* (π_1, π_2) and *Elaboration* (π_2, π_3) ; an annotation A_2 , containing these two relations plus the relation *Result* (π_1, π_3) . These annotations are equivalent: A_2 can be deduced from A_1 , because *Result* (π_1, π_3) is implicitly contained in A_1 .

- (1) a. It has rained a lot today. (π_1)
b. So John cooked. (π_2)
c. He made a pie. (π_3)

Even though these two annotations are equivalent, if they are compared without deducing the implicit relation, in the case of evaluating a system the score will be different from 1, and in the case where they are to be merged, it's not clear whether the implicit relation must be taken in the gold annotation or

not. This example shows that all the implicit information contained in the discourse structures must be deduced in order to improve their comparison. Given a discourse structure associated to a text, we must be able to calculate, using inference rules, the *discourse closure* of the structure, i.e. all discourse relations that can be inferred from the relations explicitly annotated. To do that, discourse relations inference rules are necessary, based on the semantic constraints that relate different discourse functions. For example, the calculation of the discourse closure of the annotation A_1 requires the following rule: $Result(\pi_1, \pi_2) \wedge Elaboration(\pi_2, \pi_3) \rightarrow Result(\pi_1, \pi_3)$. Despite the interest of using such constraints for a better comparison of discourse structures, interactions between relations and equivalences between structures remain little studied in discourse theories.¹ A very small number of discourse relations inference rules have been built. In many cases, we don't know if a typical structure such as $R_1(\alpha, \beta) \wedge R_2(\beta, \gamma)$ ² contains implicit information(s), i.e. implicit relation(s), and if so, which one(s). We therefore propose to study and define discourse relations inference rules, to be integrated into a general framework akin to the algebra of temporal relations put forth by Allen (1983).

Several theories and formalisms aim at representing discourse structures. In RST (*Rhetorical Structure Theory*, Mann and Thompson, 1988), discourse structures are represented by binary trees (Marcu, 1996), whose leaf nodes are discourse units and internal nodes are discourse relations. On the other hand, in SDRT (*Segmented Discourse Representation Theory*, Asher and Lascarides, 2003), discourse structures are represented by graphs, whose nodes are discourse units and edges are discourse relations. The set of discourse relations also vary according to the different theories and formalisms, as well as the way they are defined. There exists multiplying approaches, such as RST (78 relations in the RST corpus (Carlson and Marcu, 2001)), which build extended lists of relations, and reductionist approaches, such as the one of Grosz and Sidner (1986), which propose to distinguish only two structural relations (dominates and satisfaction-precedence). Nevertheless, discourse relations can be classified into 4 main groups (Halliday and Hasan, 1976): additive relations, temporal relations, causal relations and adversative relations – corresponding respectively to EXPANSION, TEMPORAL, CONTINGENCY and COMPARISON in the hierarchy of sense tags from the PDTB (The PDTB Research Group, 2008). Besides, there is a consensus on a number of relations, such as *Narration* or *Result*. Discourse relations can generally be matched from one theory to another: for example, the relation *Background* from SDRT (Asher and Lascarides, 2003) overlaps the relations *Background* and *Circumstance* from RST (Mann and Thompson, 1988). We chose to build the inference rules within the framework of SDRT, because this theory makes explicit the semantic constraints established by discourse re-

¹ Some inference rules are defined within SDRT (Asher and Lascarides, 2003). For example, there is a transitivity rule for *Elaboration*: $Elaboration(\pi_1, \pi_2) \wedge Elaboration(\pi_2, \pi_3) \rightarrow Elaboration(\pi_1, \pi_3)$ (where (π_1) , (π_2) and (π_3) are three discourse segments).

² R_1 and R_2 represent discourse relations. (α) , (β) and (γ) represent discourse units.

lations on their arguments, and these constraints constitute a starting point for the study of inference rules. Besides, with respect to the definition of discourse relations, this theory is at a level of granularity intermediary between reductionist approaches and multiplying ones. In our work, we focus on discourse relations that can be matched with a relation (or a group of relations) in RST (Mann and Thompson, 1988; Carlson and Marcu, 2001) and in the PDTB. For example: *Narration* (SEQUENCE in RST, TEMPORAL:Asynchronous:precedence in the PDTB), *Background* (BACKGROUND and CIRCUMSTANCE in RST, TEMPORAL:Synchronous in the PDTB), *Result* (RESULT in the RST corpus, CONTINGENCY:Cause:result in the PDTB), etc.

Inference rules have been defined for temporal relations: temporal constraints have been used as inference rules for completing and evaluating temporal graphs (Setzer et al., 2003). For example, if a temporal annotation specifies that the event e_1 happens before e_2 and the event e_2 happens before e_3 , then it implicitly specifies that the event e_1 happens before e_3 . With respect to temporal relations, Allen (1983) defines a complete temporal algebra, containing rules of the form: $r_1(A, B) \wedge r_2(B, C) \rightarrow r_3(A, C)$ ³. In many cases, more than one relation r_3 can be deduced between A and C . For example: $overlaps(A, B) \wedge overlaps(B, C) \rightarrow before(A, C) \vee overlaps(A, C) \vee meets(A, C)$. Building inference rules for discourse relations is not as easy, since their semantic effects are not as clear as constraints established by temporal relations. To build a discourse relations algebra, two forms of rules (at least) seem necessary. For a discourse with three consecutive discourse units (α), (β) and (γ), there are at least potentially two structures (see figure 1) in which a relation remains undefined (dashed lines in the figure): in the first structure, this is the relation between (α) and (γ) and in the second one, the relation between (β) and (γ). Considering those two structures, we propose two forms of inference rules (see figure 1).⁴ The deduction can be a disjunction of relations, as in Allen’s temporal algebra. For completion of the algebra, we add an artificial relation *None* in the set of discourse relations, expressing the fact that two discourse segments are unrelated by a discourse relation, and which is exclusive of other relations.

2 Building Rules of Inference for Discourse Relations

We detail in this section the methodology we adopt for building discourse relations inference rules. For each premise of a rule, the study is based on the semantic effects of the relations contained in the premise, and the analysis of both intuitively constructed and automatically extracted data.

³ A , B et C represent temporal intervals. r_1 , r_2 and r_3 represent temporal relations.

⁴ If the relations were temporal ones, the premise $R_x(\alpha, \beta) \wedge R_z(\alpha, \gamma)$ would be equivalent to $R_x^{-1}(\beta, \alpha) \wedge R_z(\alpha, \gamma)$, where R_x^{-1} is the inverse relation of R_x . In Allen’s algebra, the premises of rules are all of the form: $r_1(A, B) \wedge r_2(B, C)$. With respect to discourse relations, these two cases must be distinguished because not all discourse relations have an inverse relation, and the ordering of segments in a discourse have consequences on its structure and its interpretation.



Fig. 1. Representation of the two forms of rules: inference of R_z or inference of R_y (where (α) , (β) and (γ) are contiguous discourse units)

Formulating Candidate Rules At first, we examine the constraints established by the relations contained in the premise of rule under consideration (either R_x and R_y or R_x and R_z), using their theoretical definition, in order to characterize the possible links between the two segments whose relation is unknown (either R_z or R_y). The constraints established by the relations of the premise allow us to predict the inference, at least theoretically. For example, the premise $Narration(\alpha, \beta) \wedge Explanation(\beta, \gamma)$ has the following temporal effects: $e_\alpha < e_\beta$ and $Init(e_\beta) < Init(e_\gamma)$. It leaves several possibilities open about the temporal links between e_α and e_γ and leads to the formulation of a possible rule: $Narration(\alpha, \beta) \wedge Explanation(\beta, \gamma) \rightarrow Narration(\alpha, \gamma) \vee Background(\alpha, \gamma) \vee Flashback(\alpha, \gamma)$. These three possibilities are illustrated by the discourses in (2), (3) and (4).

- (2) Peter lived in Paris. *Then* he moved to South *because* he was offered a post in Marseille. $\rightarrow Narration(\alpha, \gamma)$
- (3) Pierre spent two weeks in London, *then* he went to Greece *because* it was raining all the time. $\rightarrow Background(\alpha, \gamma)$
- (4) Peter went shopping. *then* he called his sister, *because* she left him a message yesterday. $\rightarrow Flashback(\alpha, \gamma)$

Studying the constraints established by the relations of the premise also allows for the exclusion of specific relations from possible inferences. For example, for the premise $Narration(\alpha, \beta) \wedge Narration(\beta, \gamma)$, the inference of $Flashback(\alpha, \gamma)$ is impossible: given that the eventuality described in (α) happens before the one in (β) and the eventuality described in (β) happens before the one in (γ) , the eventuality described in (α) cannot happen after the one in (γ) .

The study of inference rules also must get through an introspective work, i.e. building and analysing discourses containing the premise under consideration, trying to cover the different realizations of the involved discourse relations. This work allows for the verification of the inference hypotheses resulting from the study of semantic effects of discourse relations.

Tests for Identification of Inferences The analysis of discourses containing the premise can be helped by the use of two tests: the insertion of a connective and a discourse reorganization. These tests are used for checking the presence of a specific relation R_z between (α) and (γ) in the case of a premise $R_x(\alpha, \beta) \wedge R_y(\beta, \gamma)$. The first test is the following: after inserting in the segment (γ) a discourse connective marking R_z , if the discourse remains coherent and its

interpretation is unchanged, then $R_z(\alpha, \gamma)$ holds. The second test requires the reorganization of the discourse, by inverting the positions of (β) and (γ) . The connective marking $R_y(\beta, \gamma)$ is replaced by a connective marking $R_y^{-1}(\gamma, \beta)$ (if the inverse relation R_y^{-1} exists). In the resulting discourse, if a connective marking $R_z(\alpha, \gamma)$ can be inserted, without changing the coherence and the original interpretation of the discourse, then the relation $R_z(\alpha, \gamma)$ holds in the original discourse.

In (5) we can see two examples of the insertion test, where we check the presence of the relation $Result(\alpha, \gamma)$ within the study of the premise $Result(\alpha, \beta) \wedge Explanation(\beta, \gamma)$, by inserting the connective *therefore*: in (5-i) the insertion is possible, and in (5-ii) it is not. The two discourses are reorganized in (6), by inverting the positions of (β) and (γ) and inserting the connective *so*: the reorganization is possible in (6-i), but it is not in (6-ii).

- (5) α . The electricity has been restored this morning.
 β . The inhabitants of the building are relieved,
 γ . (i) *because as a result* they have been able to return to their homes.
(ii) *because (# as a result)* they need heating.
- (6) α . The electricity has been restored this morning.
 γ . (i) *So* the inhabitants of the building have been able to return to their homes.
(ii) *# So* the inhabitants of the building need heating.
 β . *Therefore* they are relieved.

Annotation After identifying the possible inferences for a premise, we proceed to a systematic annotation of automatically extracted discourses which contain the premise. For gathering a corpus of discourses containing a specific premise, we developed a tool that identifies the presence of discourse relations by detecting discourse connectives marking them. The extraction is done on the French journalistic corpus Est Républicain and the corpus Europarl, in which syntactic dependencies were annotated beforehand by the BONSAI parser (Candito et al., 2009), using LEXCONN, a lexicon of French discourse connectives (Roze et al., 2010) containing 330 connectives collected with their syntactic category and the discourse relation(s) they express. For extracting occurrences of a premise $R_x(\alpha, \beta) \wedge R_y(\beta, \gamma)$, the tool detects the following contexts in the corpus: $p_1 p_2 [conn_x] p_3 [conn_y]$, where the connective $conn_x$ appears in the proposition p_2 and is a marker for R_x , and $conn_y$ appears in the proposition p_3 and is a marker for R_y . In the collected discourses, either $conn_y$ is in the same sentence as $conn_x$ or in the next sentence. This method only allows for the extraction of discourses where the relations are explicitly marked, but it permits to gather a greater number of examples than in annotated corpora (for example, 2452 discourses were extracted for the premise $Explanation(\alpha, \beta) \wedge Result(\beta, \gamma)$).

The annotation of discourses contained in the gathered corpus consists in: verifying the presence of the two relations of the premise; if the premise is present

in the discourse, using the tests presented in the last section in order to identify and annotate the inferred relation. This annotation allows for the verification of the hypotheses formulated about possible inferences and in the case where a premise gives rise to more than one inference, it is also possible to estimate the frequencies of each result. Within the study of the premise of rule $Result(\alpha, \beta) \wedge Contrast(\beta, \gamma)$, we annotated 171 discourses containing the premise (Table 2). The study of the premise allows the formulation of the rule: $Result(\alpha, \beta) \wedge Contrast(\beta, \gamma) \rightarrow Contrast(\alpha, \gamma) \vee Result(\alpha, \gamma) \vee None(\alpha, \gamma)$. The results of the annotation show that in most cases, the relation *Contrast* is inferred (75%).

Inferred Relation	Percentage	Number
<i>Contrast</i>	75	128
<i>Result</i>	15	25
<i>None</i>	7	12
Other	6	10
Total	103	170

Table 1. Percentage of inferred relations between (α) and (γ) in extracted discourses containing the premise $Result(\alpha, \beta) \wedge Contrast(\beta, \gamma)$

3 Conclusion and Future Work

We proposed a methodology for building discourse relations inference rules, to be integrated into an algebra of these relations. The main objective of building such an algebra is to improve the comparison of discourse structures within the evaluation of discourse annotations and the creation of a gold-standard corpus. The inference rules can also help detecting inconsistencies in discourse structures, in order to improve human or machine annotation. The premises of rules already studied lead to the formulation of inference rules, established by the theoretical definition of discourse relations, manually constructed data and extracted data. By manually annotating discourses, we also compute inference probabilities.

The construction of more inference rules is under way. As for rules that give rise to the deduction of a disjunction of relations, we aim to determine which linguistic features help to predict the inferred relation, based on a linguistic study and an investigation of annotated data by statistical methods. We will also seek to establish generalizations over rules, and attempt to determine whether the type of relation (coordinating or subordinating) has an impact on the inference, and whether relations sharing some semantic effects have similar behavior in inference rules. Another issue is that our study is somewhat limited to discourses where (α) , (β) and (γ) are elementary discourse units, so we have to address the following question: are the inference rules still valid when (α) , (β) and (γ) are complex discourse units? We are already trying to answer that question within the study of some rules.

Bibliography

- James Allen. Maintaining knowledge about temporal intervals. In *Communications of the ACM*. ACM Press, 1983.
- Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.
- Marie Candito, Benoit Crabbé, Pascal Denis, and François Guérin. Analyse syntaxique du français : des constituants aux dépendances. In *Proceedings of TALN'09*, Senlis, France, 2009.
- Lynn Carlson and Daniel Marcu. Discourse Tagging Manual. Technical Report ISI-TR-545, July 2001.
- Barbara Grosz and Candace Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204, 1986.
- M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. London: Longman, 1976.
- William Mann and Sandra Thompson. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8:243–281, 1988.
- Daniel Marcu. Building up rhetorical structure trees. In *Proceedings of 13th National Conference on Artificial Intelligence*, volume 2, pages 1069–1074, Portland, Oregon, 1996.
- Charlotte Roze, Laurence Danlos, and Philippe Muller. LEXCONN: a French Lexicon of Discourse Connectives. In *Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010)*, Moissac, France, 2010.
- Andrea Setzer, R. Gaizauskas, and M. Hepple. Using semantic inferences for temporal annotation comparison. In *Proceedings of the Fourth International Workshop on Inference in Computational Semantics (ICoS-4)*, 2003.
- The PDTB Research Group. The Penn Discourse Treebank 2.0. Annotation Manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA, 2008.