# Testing SDRT's Right Frontier

**Stergos D. Afantenos** and **Nicholas Asher**
Institut de recherche en informatique de Toulouse (IRIT),
CNRS, Université Paul Sabatier
{stergos.afantenos, nicholas.asher}@irit.fr

## Abstract

The Right Frontier Constraint (RFC), as a constraint on the attachment of new constituents to an existing discourse structure, has important implications for the interpretation of anaphoric elements in discourse and for Machine Learning (ML) approaches to learning discourse structures. In this paper we provide strong empirical support for SDRT's version of RFC. The analysis of about 100 doubly annotated documents by five different naive annotators shows that SDRT's RFC is respected about 95% of the time. The qualitative analysis of presumed violations that we have performed shows that they are either click-errors or structural misconceptions.

## 1 Introduction

A cognitively plausible way to view the construction of a discourse structure for a text is an incremental one. Interpreters integrate discourse constituent $n$ into the antecedently constructed discourse structure $D$ for constituents 1 to $n-1$ by linking $n$ to some constituent in $D$ with a discourse relation. SDRT's Right Frontier Constraint (RFC) (Asher, 1993; Asher and Lascarides, 2003) says that a new constituent $n$ cannot attach to an arbitrary node in $D$. Instead it must attach to either the last node entered into the graph or one of the nodes that dominate this last node. Assuming that the last node is usually found on the right of the structure, this means that the nodes available for attachment occur on the *right frontier* (RF) of the discourse *graph* or SDRS.

Researchers working in different theoretical paradigms have adopted some form of this constraint. Polanyi (1985; 1988) originally proposed the RFC as a constraint on antecedents to anaphoric pronouns. SDRT generalizes this to a condition on all anaphoric elements. As the attachment of new information to a contextually given discourse graph in SDRT involves the resolution of an anaphoric dependency, RFC furnishes a constraint on the attachment problem. (Webber, 1988; Mann and Thompson, 1987; 1988) have also adopted versions of this constraint. But there are important differences. While SDRT and RST both take RFC as a constraint on all discourse attachments (in DLTAG, in contrast, anaphoric discourse particles are not limited to finding an antecedent on the RF), SDRT's notion of RF is substantially different from that of RST's or Polanyi's, because SDRT's notion of a RF depends on a 2-dimensional discourse graph built from *coordinating* and *subordinating* discourse relations. Defining RFC with respect to SDRT's 2-dimensional graphs allows the RF to contain discourse constituents that do not include the last constituent entered into the graph (in contrast to RST). SDRT also allows for multiple attachments of a constituent to the RFC.

SDRT's RFC has important implications for the interpretation of various types of anaphoric elements: tense (Lascarides and Asher, 1993), ellipsis (Hardt et al., 2001; Hardt and Romero, 2004; Asher, 2007), as well as pronouns referring to individuals and abstract entities (Asher, 1993; Asher and Lascarides, 2003). The RFC, we believe, will also benefit ML approaches to learning discourse structures, as a constraint limiting the search space for possible discourse attachments. Despite its importance, SDRT's RFC has never been empirically validated, however. We present evidence in this paper providing strong empirical support for SDRT's version of the constraint. We have chosen to study SDRT's notion of a RF, because of SDRT's greater expressive power over RST (Danlos, 2008), the greater generality of SDRT's defi-



*Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1–9,
Beijing, August 2010

nition of RFC, and because of SDRT's greater theoretical reliance on the constraint for making semantic predictions. SDRT also makes theoretically clear why the RFC should apply to discourse relation attachment, since it treats discourse structure construction as a dynamic process in which all discourse relations are essentially anaphors. The analysis of about 100 doubly annotated documents by five different naive annotators shows that this constraint, as defined in SDRT, is respected about 95% of the time. The qualitative analysis of the presumed violations that we have performed shows that they are either click-errors or structural misconceptions by the annotators.

Below, we give a formal definition of SDRT's RFC; section 3 explains our annotation procedure. Details of the statistical analysis we have performed are given in section 4, and a qualitative analysis is provided in section 5. Finally, section 6 presents the implications of the empirical study for ML techniques for the extraction of discourse structures while sections 7 and 8 present the related work and conclusions.

## 2 The Right Frontier Constraint in SDRT

In SDRT, a discourse structure or SDRS (Segmented Discourse Representation Structure) is a tuple $< A, \mathcal{F}, \text{LAST} >$, where $A$ is the set of labels representing the discourse constituents of the structure, $\text{LAST} \in A$ the last introduced label and $\mathcal{F}$ a function which assigns each member of $A$ a well-formed formula of the SDRS language (defined (Asher and Lascarides, 2003, p 138)). SDRSs correspond to $\lambda$ expressions with a continuation style semantics. SDRT distinguishes coordinating and subordinating discourse relations using a variety of linguistic tests (Asher and Vieu, 2005),[1] and isolates structural relations (Parallel and Contrast) based on their semantics.

The RF is the set of available attachment points to which a new utterance can be attached. What this set includes depends on the discourse relation used to make the attachment. Here is the definition from (Asher and Lascarides, 2003, p 148).

> Suppose that a constituent $\beta$ is to be attached to a constituent in the SDRS with a discourse relation other than Parallel or Contrast. Then the available attachment points for $\beta$ are:
>
> 1. The label $\alpha = \text{LAST}$;
> 2. Any label $\gamma$ such that:
>    (a) *i-outscopes*$(\gamma, \alpha)$ (*i.e.* $R(\delta, \alpha)$ or $R(\alpha, \delta)$ is a conjunct in $\mathcal{F}(\gamma)$ for some $R$ and some $\delta$); or
>    (b) $R(\gamma, \alpha)$ is a conjunct in $\mathcal{F}(\lambda)$ for some label $\lambda$, where $R$ is a subordinating discourse relation.
>    We gloss this as $\alpha < \gamma$.
> 3. Transitive Closure:
>    Any label $\gamma$ that dominates $\alpha$ through a sequence of labels $\gamma_1, \gamma_2, \ldots \gamma_n$ such that $\alpha < \gamma_1 < \gamma_2 < \ldots \gamma_n < \gamma$

We can represent an SDRS as a graph $\mathcal{G}$, whose nodes are the labels of the SDRSs constituents and whose typed arcs represent the relations between them. The nodes available for attachment of a new element $\beta$ in $\mathcal{G}$ are the last introduced node LAST and any other node dominating LAST, where the notion of domination should be understood as the transitive closure over the arrows given by *subordinating* relations or those holding between a complex segment and its parts. Subordinating relations like *Elaboration* extend the vertical dimension of the graph, whereas coordinating relations like *Narration* expand the structure horizontally. The graph of every SDRS has a unique top label for the whole structure or formula; however, there may be multiple $<$ paths defined within a given SDRS, allowing for multiple parents, in the terminology of (Wolf and Gibson, 2006). Furthermore, SDRT allows for multiple arcs between constituents and attachments to multiple constituents on the RFC, making for a very rich structure.

SDRT's RFC is restricted to non-structural relations, because structural relations postulate a partial isomorphism from the discourse structure of the second constituent to the discourse structure of the first, which provides its own attachment possibilities for subconstituents of the two related structures (Asher, 1993). Sometimes such parallelism or contrast, also known as *discourse subordination* (Asher, 1993), can be enforced in a long

[1] The subordinating relations of SDRT are currently: Elaboration (a relation defined in terms of the main eventualities of the related constituents), Entity-Elaboration (E-Elab(a,b) iff b says more about an entity mentioned in a that is not the main eventuality of a) Comment, Flashback (the reverse of Narration), Background, Goal (intentional explanation), Explanation, and Attribution. The coordinating relations are: Narration, Contrast, Result, Parallel, Continuation, Alternation, and Conditional, all defined in Asher and Lascarides (2003).

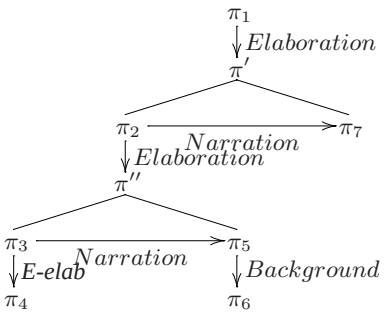
distance way by repeating the same wording in the two constituents.

RFC has the name it does because the segments that belong on this set (the $\gamma$s in the above definition) are typically nodes on a discourse graph which are geometrically placed at the RF of the graph. Consider the following example embellished from Asher and Lascarides (2003):

(1) ($\pi_1$) John had a great evening last night. ($\pi_2$) He first had a great meal at Michel Sarran. ($\pi_3$) He ate profiterolles de foie gras, ($\pi_4$) which is a specialty of the chef. ($\pi_5$) He had the lobster, ($\pi_6$) which he had been dreaming about for weeks. ($\pi_7$) He then went out to a several swank bars.

The graph of the SDRS for 1 looks like this:

(2)

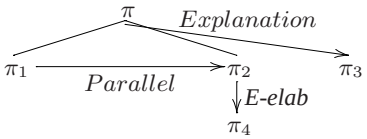



where $\pi'$ and $\pi''$ represent complex segments. Given that the last introduced utterance is represented by the node $\pi_7$, the set of nodes that are on the RF are $\pi_7$ (LAST), $\pi'$ (the complex segment that includes $\pi_7$) and $\pi_1$ (connected via a subordinating relation to $\pi'$). All those nodes are geometrically placed at the RF of the graph.

SDRT's notion of a RF is more general than RST's or DLTAG's. First, SDRSs can have complex constituents with multiple elements linked by coordinate relations that serve as arguments to other relations, thus permitting instances of *shared structure* that are difficult to capture in a pure tree notation (Lee et al., 2008). In addition, in RST the RF picks out the *adjacent* constituents, LAST and complex segments including LAST. Contrary to RST, SDRT, as it uses 2-dimensional graphs, predicts that an available attachment point for $\pi_7$ is the non local and non adjacent $\pi_2$, which is distinct from the complex constituent consisting of $\pi_2$ to $\pi_6$.[2] This difference is crucial to the interpretation of the Narration: Narration claims a sequence of two events; making the complex constituent (essentially a sub-SDRS) an argument of Narration, as RST does, makes it difficult to recover such an interpretation. Danlos's (2008) interpretation of the Nuclearity Principle provides an interpretation of the Narration([2-4],5) that is equivalent to the SDRS graph above.[3] But even an optional Nuclearlity Principle interpretation won't help with discourse structures like (2) where the backgrounding material in $\pi_4$ and the commentary in $\pi_6$ do not and cannot figure as part of the Elaboration for semantic reasons. In our corpus described below, over 20% of the attachments were non adjacent; *i.e.* the attachment point for the new material did not include LAST.

A further difference between SDRT and other theories is that, as SDRT's RFC is applied recursively over complex segments within a given SDRS, many more attachment points are available in SDRT. E.g., consider the SDRS for this example, adapted from (Wolf and Gibson, 2006):

(3) ($\pi_1$) Mary wanted garlic and thyme. ($\pi_2$) She also needed basil. ($\pi_3$) The recipe called for them. ($\pi_4$) The basil would be hard to come by this time of year.




Because $\pi$ is the complex segment consisting of $\pi_1$ and $\pi_2$, attachment to $\pi$ with a subordinating discourse relation permits attachment $\pi$'s open constituents as well.[4]

## 3 Annotated Corpus

Our corpus comes from the discourse structure annotation project ANNODIS[5] which represents an on going effort to build a discourse graph bank for French texts with the two-fold goal of testing various theoretical proposals about discourse

[2]The 2-dimensionality of SDRSs also allows us to represent many examples with Elaboration that involve crossing dependencies in Wolf and Gibson's (2006) representation without violation of the RFC.

[3]Baldridge et al. (2007), however, show that the Nuclearity Principle does not always hold.

[4]This part of the RFC was not used in (Asher and Lascarides, 2003).

[5]http://w3.erss.univ-tlse2.fr/annodis

structure and providing a seed corpus for learning discourse structures using ML techniques. ANNODIS's annotation manual provides detailed instructions about the segmentation of a text into Elementary Discourse Units (EDUs). EDUs correspond often to clauses but are also introduced by frame adverbials,[6] appositive elements, correlative constructions (*[the more you work,] [the more you earn]*), interjections and discourse markers within coordinated VPs *[John denied the charges] [but then later admitted his guilt]*. Appositive elements often introduce *embedded* EDUs; e.g., *[Jim Powers, [President of the University of Texas at Austin], resigned today.]*, which makes our segmentation more fine-grained than Wolf and Gibson's (2006) or annotation schemes for RST or the PDTB.

The manual also details the meaning of discourse relations but says nothing about the structural postulates of SDRT. For example, there is no mention of the RFC in the manual and very little about hierarchical structure. Subjects were told to put whatever discourse relations from our list above between constituents they felt were appropriate. They were also told that they could group constituents together whenever they felt that as a whole they jointly formed the term of a discourse relation. We purposely avoided making the manual too restrictive, because one of our goals was to examine how well SDRT predicts the discourse structure of subjects who have little knowledge of discourse theories.

In total 5 subjects with little to no knowledge of discourse theories that use RFC participated in the annotation campaign. Three were undergraduate linguistics students and two were graduate linguistics students studying different areas. The 3 undergraduates benefitted from a completed and revised annotation manual. The two graduate students did their annotations while the annotation manual was undergoing revisions. All in all, our annotators doubly annotated about 100 French newspaper texts and *Wikipedia* articles. Subjects first segmented each text into EDUs, and then they were paired off and compared their segmentations, resolving conflicts on their own or via a supervisor. The annotation of the discourse relations was performed by each subject working in isolation. ANNODIS provided a new state of the art tool, GLOZZ, for discourse annotation for the three undergraduates. With GLOZZ annotators could isolate sections of text corresponding to several EDUs, and insert relations between selected constituents using the mouse. Though it did portray relations selected as lines between parts of the text, GLOZZ did not provide a discourse graph or SDRS as part of its graphical interface. The representation often yielded a dense number of lines between segments that annotators and evaluators found hard to read. The inadequate interline spacing in GLOZZ also contributed to certain number of click errors that we detail below in the paper. The statistics on the number of documents, EDUs and relations provided by each annotator are in table 1.

| *annotator* | *# Docs* | *# EDUs* | *# Relations* |
|---|---|---|---|
| ***undergrad 1*** | 27 | 1342 | 1216 |
| ***undergrad 2*** | 31 | 1378 | 1302 |
| ***undergrad 3*** | 31 | 1376 | 1173 |
| ***grad 1*** | 47 | 1387 | 1390 |
| ***grad 2*** | 48 | 1314 | 1321 |

Table 1: Statistics on documents, EDUs and Relations.

## 4 Experiments and Results

Using ANNODIS's annotated corpus, we checked for all EDUs $\pi$, whether $\pi$ was attached to a constituent in the SDRS built from the previous EDUs in a way that violated the RFC. Given a discourse as a series of EDUs $\pi_1, \pi_2, \ldots, \pi_n$, we constructed for each $\pi_i$ the corresponding sub-graph and calculated the set of nodes on the RF of this subgraph. We then checked whether the EDU $\pi_{i+1}$ was attached to a node that was found in this set. We also checked whether any newly created complex segment was attached to a node on the RF of this sub-graph.

### 4.1 Calculating the Nodes at the RF

To calculate the nodes on the RF, we slightly extended the annotated graphs, in order to add im-

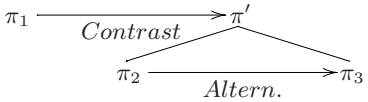
[6] Frame adverbials are sentence initial adverbial phrases that can either be temporal, spatial or "topical" (*in Chemistry*).

plied relations left out by the annotators.[7]

**Disconnected Graphs** While checking the RFC for the attachment of a node $n$, the SDRS graph at this point might consist of 2 or more disjoint subgraphs which get connected together at a later point. Because we did not want to decide which way these graphs should be connected, we defined a right frontier for each one using its own LAST. We then calculated the RF for each one of them and set the set of available nodes to be those in the union of the RFs of the disjoint subgraphs. If the subgraphs were not connected at the end of the incremental process in a way that conformed to RFC, we counted this as a violation. Annotators did not always provide us with a connected graph.

**Postponed Decisions** SDRT allows for the attachment not only of EDUs but also of subgraphs to an available node in the contextually given SDRS. For instance, in the following example, the intended meaning is given by the graph in which the Contrast is between the first label and the complex constituent composed of the disjunction of $\pi_2$ and $\pi_3$.

($\pi_1$) Bill doesn't like sports. ($\pi_2$) But Sam does.
($\pi_3$) Or John does.




Naive annotators attached subgraphs instead of EDUs to the RF with some regularity (around 2%). This means that an EDU $\pi_{i+1}$ could be attached to a node that was not present in the subgraph produced by $\pi_1, \ldots, \pi_i$. There were two main reasons for this: (1) $\pi_{i+1}$ came from a syntactically fronted clause, a parenthetical or apposition in a sentence whose main clause produced $\pi_{i+2}$ and $\pi_{i+1}$ was attached to $\pi_{i+2}$; (2) $\pi_{i+1}$ was attached to a complex segment $[\ldots, \pi_{i+1}, \ldots, \pi_{i+k}, \ldots]$ which was not yet introduced in the subgraph.

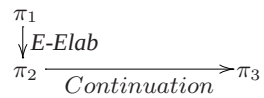
Since the nodes to which $\pi_{i+1}$ is attached in such cases are not present in the graph, *by definition* they are not in the RF and they could be counted as violations. Nonetheless, if the nodes which connect nodes like $\pi_{i+1}$ eventually link up to the incrementally built SDRS in the right way, $\pi_{i+1}$ might eventually end up linked to something on the RF. For this reason, we postponed the decision on nodes like $\pi_{i+1}$ until the nodes to which they are attached were explicitly introduced in the SDRS.

**The Coherence of Complex Segments** In an SDRS, several EDUs may combine to form a complex segment $\alpha$ that serves as a term for a discourse relation $R$. The interpretation of the SDRS implies that all of $\alpha$'s constituents contribute to the rhetorical function specified by $R$. This implies that the coordinating relation *Continuation* holds between the EDUs inside $\alpha$, unless there is some other relation between them that is incompatible with Continuation (like a subordinating relation). Continuations are often used in SDRT (Asher, 1993; Asher and Lascarides, 2003). During the annotation procedure, our subjects did not always explicitly link the EDUs within a complex segment. In order to enforce the coherence of those complex segments we added *Continuation* relations between the constituents of a complex segment *unless* there was already another path between those constituents.

**Expanding Continuations** Consider the following discourse:

(4) [John, [who owns a chain of restaurants]$_{\pi_2}$, [and is a director of a local charity organization,]$_{\pi_3}$ wanted to sell his yacht.]$_{\pi_1}$ [He couldn't afford it anymore.]$_{\pi_4}$

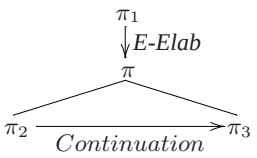
Annotators sometimes produced the following SDRT graph for the first three EDUs of this discourse:

(5)



In this case the only open node is $\pi_3$ due to the coordinating relation *Continuation*. Nonetheless, $\pi_4$ should be attached to $\pi_1$, without violating the RFC. Indeed, SDRT's definition of the Continuation relation enforces that if we have $R(\pi_1, \pi_2)$ and Continuation$(\pi_2, \pi_3)$ then we actually have the complex segment $[\pi_2, \pi_3]$ with $R(\pi_1, [\pi_2, \pi_3])$. So there is in fact a missing complex segment in (5). The proper SDRS graph of (4) is:

[7] In similar work on TimeML annotations, Setzer et al. (2003; Muller and Raymonet (2005) add implied relations to annotated, temporal graphs.

(6)




which makes $\pi_1$ an available attachment site for $\pi_4$. Such implied constituents have been added to the SDRS graphs.

**Factoring** Related to the operation of Expansion, SDRT's definition of Continuation and various subordinating relations also requires that if we have $R(a, [\pi_1, \pi_2, \ldots, \pi_n])$ where $[\pi_1, \pi_2, \ldots, \pi_n]$ is a complex segment with $\pi_1, \ldots \pi_n$ linked by Continuation and $R$ is Elaboration, Entity-Elaboration, Frame, Attribution, or Commentary, then we also have $R(a, \pi_i)$ for each $i$. We added these relations when they were missing.

### 4.2 Results

With the operations just described, we added several inferred relations to the graph. We then calculated statistics concerning the percentage of attachments for which the RFC is respected using the following formula:

$$RFC_{\text{EDU}} = \frac{\texttt{\# EDUs attached to the RF}}{\texttt{\# EDUs in total}}$$

As we explained, an EDU can be attached to an SDRT graph directly by itself or indirectly as part of a bigger complex segment. In order to calculate the nominator we determine first whether an EDU directly attaches to the graph's RF, and if that fails we determine whether it is part of a larger complex segment which is attached to the graph's RF. The results obtained are shown in the first two columns of table 2. The RFC is respected by at least some attachment decision 95% of the time—i.e., 95% of the EDUs get attached to another node that is found on the RF. The breakdown across our annotators is given in table 2.

SDRT allows for multiple attachments of an EDU to various nodes in an SDRS; *e.g.* while an EDU may be attached via one relation to a node on the RF, it may be attached to another node off the RF. To take account of all the attachments for a given EDU, we need another way of measuring the percentage of attachments that respects the RFC. So we counted the ways each EDU is related to a node in the SDRS for the previous text and then divided the number of attachment decisions that respect the RFC by the total number of attachment decisions—*i.e.* :

$$RFC_r = \frac{\texttt{\# RF attachment decisions}}{\texttt{\# Total attachment decisions}}$$

.

| *annotator* | $RFC_{\text{EDU}}$ | $RFC_r$ |
|---|---|---|
| ***undergrad 1*** | 98.57% | 91.28% |
| ***undergrad 2*** | 98.12% | 94.39% |
| ***undergrad 3*** | 91.93% | 89.17% |
| ***grad 1*** | 94.38% | 86.54% |
| ***grad 2*** | 92.68% | 83.57% |
| ***Mean for all annotators*** | 95.24% | 88.91% |
| ***Mean for 3 undergrad*** | 96.17% | 91.71% |

Table 2: The % with which each annotator has respected SDRT's RFC using the EDU and attachment decision measures.

The third column of table 2 shows that having a stable annotation manual and GLOZZ improved the results across our two annotator populations, even though the annotation manual did not say anything about RFC or about the structure of the discourse graphs. Moreover, the distribution of violations of the RFC follows a power law and only 4.56% of the documents contained more than 5 violations. This is strong evidence that there is little propagation of violations.

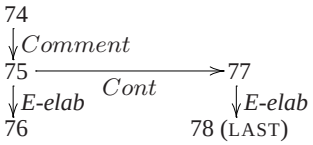
## 5 Analysis of Presumed Violations

Although 95% of EDUs attach to nodes on the RF of an SDRT graph, 5% of EDUs don't. SDRT experts performed a qualitative analysis of some of these presumed violations. In many cases, the experts judged that the presumed violations were due to click-errors: sometimes the annotators simply clicked on something that did not translate into a segment. Sometimes, the experts judged that the annotators picked the wrong segment to attach a new segment or the wrong type of relation during the construction of the SDRT graph. For example, in the graph that follows the relation between segments 74 and 75 is not a *Comment* but an *Entity-Elaboration*.

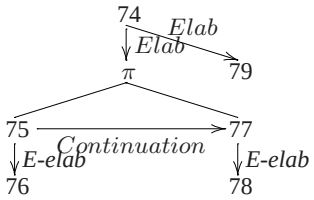
As expected, there were also *"structural"* errors, arising from a lack or a misuse of complex segments. Here is a typical example (translated from the original French):

> [Around her,]_74 [we should mention Joseph Racaille]_75 [responsible for the magnificent arrangements,]_76 [Christophe Dupouy]_77 [regular associate of Jean-Louis Murat responsible for mixing,]_78 [without forgetting her two guardian angels:]_79 [her agent Olivier Gluzman]_80 [who signed after a love at first sight,]_81 [and her husband Mokhtar]_82 [who has taken care of the family]_83

Here is the annotated structure up to EDU 78:

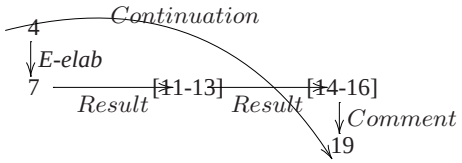



Note that the attachment of 77 to 75 is non-local and non-adjacent. The annotator then attaches EDU 79 to 75 which is blocked from the RF due to the $Continuation$ coordinating relation. By not having created a complex segment due the enumeration that includes EDUs 75 to 78, the annotator had no option but to violate the RF. Here is the proper SDRT graph for segments 74 to 79 (where the attachment of 79 to 74 is also both non-local and non-adjacent):




In this case, before the introduction of EDU 79, EDU 78 is LAST and by consequence 77, $\pi$ and 74 are on the RF. Attaching 79 to 74 is thus legitimate.

We also found more interesting examples of right frontier violations. One annotator produced a graph for a story which is about the attacks of 9/11/2001 and is too long to quote here. A simplified graph of the first part of the story is shown below. EDU 4 elaborates on the main event of the story but it is not on the RF for 19. However, 19 is the first recurrence of the complex definite description *le 11 septembre 2001* since the title and the term's definition in EDU 4.




This reuse of the full definite description could be considered a case of SDRT's discourse subordination.

## 6 RFC and distances of attachment

Our empirical study vindicates SDRT's RFC, but it also has computational implications. Using the RFC dramatically diminishes the number of attachment possibilities and thus greatly reduces the search space for any incremental discourse parsing algorithm.[8] The mean of nodes that are open on the RF at any given moment on our ANNODIS data is 16.43% of all the nodes in the graph.

Our data also allowed us to calculate the distance of attachment sites from LAST, which could be an important constraint on machine learning algorithms for constructing discourse structures. Given a pair of constituents $(\pi_i, \pi_j)$ distance is calculated either *textually* (the number of intervening EDUs between $\pi_i$ and $\pi_j$) or *topologically* (the length the shortest path between $\pi_i$ and $\pi_j$). Topological distance, however, does not take into account the fact that a textually further segment is cognitively less salient. Moreover, this measure can give the same distance to nodes that are textually far away between them due to long distance pop-ups (Asher and Lascarides, 2003). A purely textual distance, on the other hand, gives the same distance to an EDU $\pi_i$ and a complex segment $[\pi_1, \ldots, \pi_i]$ even if $\pi_1$ and $\pi_i$ are textually distant (since both have the same span end). We used a measure combining both. The distance scheme that we used assigns to each EDU its textual distance from LAST in the graph under consideration, while a complex segment of rank 1 gets a distance which is computed from the highest distance of their constituent EDUs plus 1. For a constituent $\sigma$ of rank $n$ we have:

$$Dist = Max\{\text{dist}(x) \colon x \text{ in } \sigma\} + n$$

[8] An analogous approach for search space reduction is followed by duVerle and Prendinger (2009) who use the "Principle of Sequentiality" (Marcu, 2000), though they do not say how much the search space is reduced.

The distribution of attachment follows a power law with 40% of attachments performed non-locally, that is on segments of distance 2 or more (figure 1). This implies that the distance between candidate attachment sites that are on the RF is an important feature for an ML algorithm. It is important to note at this point that following the baseline approach of always attaching on the LAST misses 40% of attachments. We also have 20.38% of the non-local, non-adjacent attachments in our annotations. So an RST parser using Marcu's (2000) adjacency constraint as do duVerle and Prendinger (2009) would miss these.

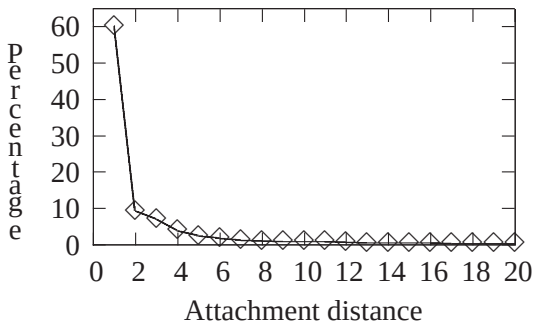



Figure 1: Distribution of attachment distance

## 7 Related Work

Several studies have shown that the RFC may be violated as an anaphoric constraint when there are other clues, content or linguistic features, that determine the antecedent. (Poesio and di Eugenio, 2001; Holler and Irmen, 2007; Asher, 2008; Prévot and Vieu, 2008), for example, show that anaphors such as definite descriptions and complex demonstratives, which often provide enough content on their own to isolate their antecedents, or pronouns in languages like German which must obey gender agreement, might remain felicitous although the discourse relations between them and their antecedents might violate the RFC. Usually there are few linguistic clues that help find the appropriate antecedent to a discourse relation, in contrast to the anaphoric expressions mentioned above. Exceptions involve stylistic devices like direct quotation that license discourse subordination. Thus, SDRT predicts that RFC violations for discourse attachments should be much more rare than those for the resolution of anaphors that provide linguistic clues about their antecedents.

As regards other empirical validation of various versions of the RFC for the attachment of discourse constituents, Wolf and Gibson (2006) show an RST-like RFC is not supported in their corpus GraphBank. Our study concurs in that some 20% of the attachments in our corpus cannot be formulated in RST.[9] On the other hand, we note that because of the 2 dimensional nature of SDRT graphs and because of the caveats introduced by structural relations and discourse subordination, the counterexamples from GraphBank against, say, RST representations do not carry over straightforwardly to SDRSs. In fact, once these factors are taken into account, the RFC violations in our corpus and in GraphBank are roughly about the same.

## 8 Conclusions

We have shown that SDRT's RFC has strong empirical support: the attachments of our 3 completely naive annotators fully comply with RFC 91.7% of the time and partially comply with it 96% of the time. As a constraint on discourse parsing SDRT's RFC, we have argued, is both empirically and computationally motivated. We have also shown that non-local attachments occur about 40% of the time, which implies that attaching directly on the LAST will not yield good results. Further, many of the non local attachments do not respect RST's adjacency constraint. We need SDRT's RFC to get the right attachment points for our corpus. We believe that empirical studies of the kind we have given here are essential to finding robust and useful features that will vastly improve discourse parsers.

[9] One other study we are aware of is Sassen and Kühnlein (2005), who show that in chat conversations, the RFC does not always hold unconditionally. Since this genre of discourse is not always coherent, it is expected that the RFC will not always hold here.

## References


Asher, N. and A. Lascarides. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK.

Asher, N. and L. Vieu. 2005. Subordinating and coordinating discourse relations. *Lingua*, 115(4):591–610.

Asher, N. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

Asher, N. 2007. A large view of semantic content. *Pragmatics and Cognition*, 15(1):17–39.

Asher, N. 2008. Troubles on the right frontier. In Benz, A. and P. Kühnlein, editors, *Constraints in Discourse*, Pragmatics and Beyond New Series, chapter 2, pages 29–52. John Benjamins Publishing Company.

Baldridge, J., N. Asher, and J. Hunter. 2007. Annotation for and robust parsing of discourse structure on unrestricted texts. *Zeitschrift fur Sprachwissenschaft*, 26:213–239.

Danlos, L. 2008. Strong generative capacity of rst, sdrt and discourse dependency dags. In Benz, A. and P. Kühnlein, editors, *Constraints in Discourse*, Pragmatics and Beyond New Series, pages 69–95. John Benjamins Publishing Company.

duVerle, D. and H. Prendinger. 2009. A novel discourse parser based on support vector machine classification. In *Proceedings of ACL*, pages 665–673, Suntec, Singapore, August.

Hardt, D. and M. Romero. 2004. Ellipsis and the structure of discourse. *Journal of Semantics*, 21:375–414, November.

Hardt, D., N. Asher, and J. Busquets. 2001. Discourse parallelism, scope and ellipsis. *Journal of Semantics*, 18:1–16.

Holler, A. and L. Irmen. 2007. Empirically assessing effects of the right frontier constraint. In *Anaphora: Analysis, Algorithms and Applications*, pages 15–27. Springer, Berlin/Heidelberg.

Lascarides, A. and N. Asher. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493.

Lee, A., R. Prasad, A. Joshi, and B. Webber. 2008. Departures from tree structures in discourse: Shared arguments in the penn discourse treebank. In *Constraints in Discourse (CID '08)*, pages 61–68.

Mann, W. and S. Thompson. 1987. Rhetorical structure theory: A framework for the analysis of texts. Technical Report ISI/RS-87-185, Information Sciences Institute, Marina del Rey, California.

Mann, W. and S. Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8(3):243–281.

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

Muller, P. and A. Raymonet. 2005. Using inference for evaluating models of temporal discourse. In *12th International Symposium on Temporal Representation and Reasoning*, pages 11–19. IEEE Computer Society Press.

Poesio, M. and B. di Eugenio. 2001. Discourse structure and anaphoric accessibility. In *Proc. of the ESSLLI Workshop on Discourse Structure and Information Structure*, August.

Polanyi, L. 1985. A theory of discourse structure and discourse coherence. In Kroeber, P. D., W. H. Eilfort, and K. L. Peterson, editors, *Papers from the General Session at the 21st Regional Meeting of the Chicago Linguistics Society*.

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

Prévot, L. and L. Vieu. 2008. The moving right frontier. In Benz, A. and P. Kühnlein, editors, *Constraints in Discourse*, Pragmatics and Beyond New Series, chapter 3, pages 53–66. John Benjamins Publishing Company.

Sassen, C. and P. Kühnlein. 2005. The right frontier constraint as conditional. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science (LNCS), pages 222–225.

Setzer, A., R. Gaizauskas, and M. Hepple. 2003. Using semantic inferences for temporal annotation comparison. In *Proceedings of the Fourth International Workshop on Inference in Computational Semantics (ICoS-4)*.

Webber, B. 1988. Title discourse deixis and discourse processing. Technical Report MS-CIS-88-75, University of Pennsylvania, Department of Computer and Information Science, September.

Wolf, F. and E. Gibson. 2006. *Coherence in Natural Language: Data Stuctures and Applications*. The MIT Press.