Natural Language Engineering Group
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
United Kingdom

# Discourse Structure and Anaphora: an Empirical Study

Massimo Poesio,[†] Barbara Di Eugenio,[‡]
Gerard Keohane[*]
[†]University of Essex, poesio@essex.ac.uk
[‡]University of Illinois at Chicago, bdieugen@cs.uic.edu
[*]TU Chemnitz, gerard.keohane@phil.tu-chemnitz.de

NLE Technical Note TN-02-02

CS Technical Report CSM-370

## University of Essex, Department of Computer Science, NLE Group

Other Technical Notes and theses from the Natural Language Engineering group are available electronically at

```
http://cswww.essex.ac.uk/Research/nledis.htm
```

# Discourse Structure and Anaphora: an Empirical Study

## Abstract

One of the main motivations for studying discourse structure is its effect on the search for the antecedents of anaphoric expressions. We tested the predictions in this regard of theories assuming that the structure of a discourse depends on its intentional structure, such as Grosz and Sidner's theory. We used a corpus of tutorial dialogues independently annotated according to Relational Discourse Analysis (RDA), a theory of discourse structure merging ideas from Grosz and Sidner's theory with proposals from Rhetorical Structure Theory (RST). Using as our metrics the accessibility of anaphoric antecedents and the reduction in ambiguity brought about by a particular theory, we found support for Moser and Moore's proposal that among the units of discourse assumed by an RST-like theory, only those expressing an intentional 'core' (in the RDA sense) should be viewed as constraining the search for antecedents; units only expressing informational relations should not introduce separate focus spaces. We also found that the best compromise between accessibility and ambiguity ('perplexity') reduction is a model in which the focus spaces associated with embedded cores and embedded contributors remain on the stack until the RDA-segment in which they occur is completed, and discuss the implications of this finding for a stack-based theory.

## 1 Introduction

One of the main motivations for studying discourse structure is its effect on the search for the antecedents of anaphoric expressions. The recent development of more reliable annotation techniques, and the increased availability of corpora annotated for discourse structure (Carletta et al. 1997; Moser et al. 1996; Marcu 1999), have made it possible to subject the claims of seminal theories about the impact of discourse structure on anaphora such as (Reichman 1985; Grosz and Sidner 1986; Fox 1987) to rigorous empirical testing. Quite a lot of this work has focused on studying the claims of theories based on Rhetorical Structure Theory (RST) (Mann and Thompson 1988)–see, e.g., (Cristea et al. 1998, 2000; Ide and Cristea 2000)).[1] Our aim in this work was to study the claims of theories that hypothesize a tight connection between the space of anaphoric antecedents and intentional structure, the best known among which being the theory proposed by Grosz and Sidner (1986).

Evaluating Grosz and Sidner's theory used to be a problem, because although it has originated a coding manual (Nakatani et al. 1995) that has been used at least once (Nakatani 1996), as far as we know there is no sizeable corpus coded accordingly. However, recent proposals concerning the mapping between rhetorical structure and intentional structure (Moser and Moore 1996b) have resulted in the development of Relational Discourse Analysis (RDA) (Moore and Pollack 1992; Moser and Moore 1996b), a theory of discourse structure merging ideas from RST with ideas from Grosz and Sidner's theory. This theory has served as the basis for a coding scheme which has been used to produce corpora containing texts annotated with their intentional structure, as well as other structural properties of the type hypothesized in RST. These corpora can be used to investigate the claims of Grosz and

[1] The classic study by Fox (1987) also used RST to analyze the structure of written texts, but Fox couldn't make use at the time of standard resources annotated in a reliable way.

Sidner's theory, as well as the relation between their view of the connection between discourse structure and anaphora and views based on RST, such as Fox's or Veins Theory.  The Sherlock corpus of tutorial dialogues collected at the University of Pittsburgh and subsequently annotated according to RDA is particularly well suited for these purposes.

In this paper, we first briefly review Grosz and Sidner's theory and RDA, and discuss how an RDA analysis can be used to analyze claims about anaphora. We then discuss how we investigated such claims: our evaluation metrics, our corpus, our annotation methods, and how we used the annotated corpus to computing the metrics. In the next section of the paper, we discuss first of all the results we obtained by assuming the simplest view of the mapping between intentional structure in the sense of RDA and focus stack operations; in particular, we discuss the consequences of the more distinctive aspect of RDA, the distinction between intentional and informational relations. We then examine look at more complex ways of mapping intentional cores into Grosz and Sidner's DSPs –in particular, how embedded segments of different types should be treated.

# 2 Background

## 2.1 Grosz and Sidner's Theory

As Grosz and Sidner's (G&S) theory is well-known, we only give a quick summary of its main ideas here. According to G&S, the structure of a discourse is determined by the intentions that the people producing it intend to convey, or DISCOURSE SEGMENT PURPOSES (DSPs).[2] In a coherent discourse, all of these DSPs are related to form an INTENTIONAL STRUCTURE by either **dominance** relations (in case a particular DSP is interpreted as contributing to the satisfaction of another intention) or **satisfaction-precedes** relations (when the satisfaction of an intention is a precondition for the satisfaction of a second one).

Anaphoric accessibility of entities in a discourse is modeled by its ATTENTIONAL STRUCTURE, which, according to Grosz and Sidner, is a stack of FOCUS SPACES. G&S propose that when a segment is open, its corresponding focus space, which includes the discourse entities introduced in that segment, is pushed onto the focus stack; when the segment is closed, the focus space is popped, and the discourse entities associated with that focus space are not accessible any more. A further hypothesis of G&S is that the pushing and popping of focus spaces on the stack reflects the intentional structure, in the sense that a new focus space is pushed on the stack whenever the discourse introduces a new DSP subordinate to the present one, and the focus space of the current is popped whenever the associated DSP is satisfied.

This claim about anaphoric accessibility was illustrated in the original paper with a few examples; however, as far as we know, it has not been empirically tested.  Part of the problem is that there are no guidelines about how to identify the DSPs in a discourse. Our purpose is therefore twofold: to test G&S's claims (with respect to a certain genre and domain), but also to use the insights gained from work on RDA to clarify the notion of DSP.

## 2.2 Relational Discourse Analysis

Relational Discourse Analysis (RDA) (Moore and Pollack 1992; Moser and Moore 1996b) is a synthesis of ideas from Grosz and Sidner's theory (Grosz and Sidner 1986) and Rhetorical Structure Theory

[2] The reason for the name is that Grosz and Sidner propose that the 'discourse segments' of discourse analysis are best seen as the portions of a discourse concerned with the satisfaction of a given intention.

(RST) (Mann and Thompson 1988). RDA inherits from Grosz and Sidner's work the idea that discourse structure is determined by intentional structure: each RDA-segment originates with an intention of the speaker. But RDA-segments are also like RST spans, in that they have additional structure, in two respects: they are generally associated with relations; and their constituents have different status.

According to RDA, all constituents of a discourse are connected by relations, of which there are two types: INFORMATIONAL relations that express a connection between facts and events 'in the world' (such as causal and temporal relations) and INTENTIONAL ones that express a discourse intention (such as evidence or concession). While a similar distinction is already made in RST between SUBJECT-MATTER and PRESENTATIONAL relations ((Mann and Thompson 1988), p. 18), in RDA the distinction has further significance it that only spans of discourse tied by intentional relations form proper RDA-SEGMENTs (in the sense that they restrict anaphoric accessibility).[3] Each such segment consists of one CORE, i.e., that constituent that most directly expresses the speaker's intention,[4] and any number of CONTRIBUTORS, the remaining constituents in the segment, each of which plays a role in serving the purpose expressed by the core (e.g., they may convey information meant to support the proposition expressed by the core). The distinction between core and contributor is of course related to the distinction between nucleus and satellite in Rhetorical Structure Theory (RST) (Mann and Thompson 1988), according to which in each "segment" ('text span,' in RST) one component should be identified as the 'main' one, and the others as secondary. However, in RST there is a distinction between nucleus and satellite for all RST subordinating relations, whereas in RDA a core and contributors are only identified if a segment purpose has been recognized –i.e., only for RDA-segments.

A distinguish feature of RDA is that in RDA-segments, each contributor is linked to the core by one intentional relation and one informational relation, as we will see in a moment. This is unlike RST, in which only a single relation can obtain between nucleus and satellite; this change was proposed in (Moore and Pollack 1992). Moore and Pollack argue that in examples like (1):

(1) a. George Bush supports big business.
b. He's sure to veto House bill 1711.

the two units can be viewed as being related by both an intentional **evidence** relation (with b as a nucleus, and a as a satellite) and an informational **volitional cause** one. Furthermore, they argued that whereas Mann and Thompson claimed that only one relation had to be chosen in such cases (which they observed), preserving both relations was in fact not only useful to avoid conflicts, but necessary, to account for the flow of inference both from an interpretation and from a generation point of view.[5]

In RDA, segment constituents may in turn be other embedded segments, or simpler functional elements. These elements may be either atomic UNITS, i.e., descriptions of domain actions and states, or CLUSTERS. Clusters are spans that only involve constituents linked by informational relations; no *core:contributor* structure exists, but they can themselves be embedded. Note, however, that when the intentional structure ends – ie, the innermost segment is identified — the text may still analyzed

[3] A similar distinction is also made in 'structured' versions of Discourse Representation Theory such as CRT (Poesio and Traum 1997), in which temporal and causal relations between events are part of the propositions expressed by speech acts, whereas a second category of relations relates the speech acts to eachother.

[4] The core may be implicit: the core of an answer, for example, often turns out to be the presupposition of the question.

[5] It might be argued that once the units of discourse are analyzed in greater depth, distinguishing between their propositional content and their force, informational relations and intentional relations may be seen as relating entities at different levels of analysis (e.g., described events vs. propositions vs. speech acts). Keep in mind however that we are primarily concerned here in RDA as a tool for corpus analysis; and from an annotation perspective, it is often a good idea not to attempt distinctions between entities such as events, propositions, and speech acts. As a result, marking both informational relations and intentional relations as relations between discourse units can be viewed as a useful way of simplifying the annotator's task.

1.1 Before troubleshooting inside the test station,
1.2 it is always best to eliminate both the UUT and TP.
2.1 Since the test package is moved frequently,
2.2 it is prone to damage.
3.1 Also, testing the test package is much easier and faster
3.2 than opening up test station drawers.

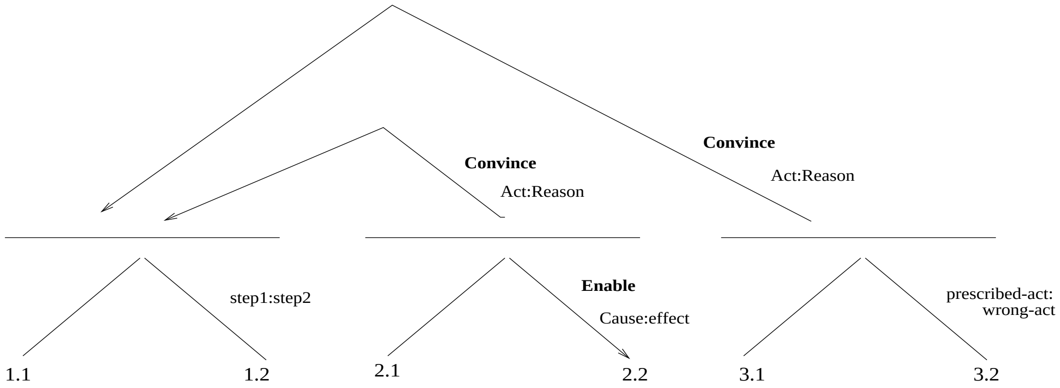

Figure 1: A tutorial excerpt and its RDA analysis

in terms of embedded clusters, but no cluster can be superordinate to an intentional segment. Unlike G&S's theory, but as in RST, RDA is based on a fixed number of relations; in particular, RDA assumes four intentional relations – **convince**, **enable**, **concede**, **joint**–and a larger set of informational relations, which is expected to be domain dependent. In the Sherlock corpus, 23 informational relations are used, of which 13 pertain to causality (they express relations between two actions, or between actions and their conditions or effects) (Moser et al. 1996).

Figure 1 shows a small excerpt from one of the dialogues in the Sherlock corpus (UUT is "Unit under test", TP is "test package"), and its RDA analysis. The analysis characterizes the text as an RDA-segment whose core spans 1.1 and 1.2. This segment has two contributors, spanning 2.1 and 2.2, and 3.1 and 3.2, respectively. (Graphically, the core is signaled as the element at the end of the arrow whose origin is the contributor; moreover, the link is marked by two relations, intentional (in bold), and informational.) In this case, the two contributors carry the same intentional and informational relations to the core, but this doesn't need to be the case. The core and the second contributor are further analyzed as informational clusters, whereas the first contributor is recognized as having its own intentional structure.[6] Clusters are marked by one informational relation, but not by intentional relations. ( In RST, the structure would presumably be the same, although no double relations would exist, and every relation would have directionality: i.e., for every relation one relatum would be considered as the nucleus, the other(s) as its satellites. We will come back to the analysis of the text according to G&S in the next section.)

[6] According to the manual used for the annotation (Moser et al. 1996), an **enable** relation holds "if the contributor [2.1] provides information intended to increase the hearer's understanding of the material presented in the core, or to increase the hearer's ability to perform the action presented in the core." (p. 6).

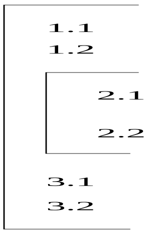

Figure 2: G&S analysis for the text in Fig. 1 on the basis of the proposed mapping of RDA-segments into G&S segments.

# 3 Questions Addressed in This Study

## 3.1 Using an RDA-annotated corpus to evaluate intentional theories

In order to use an annotation based on RDA to evaluate Grosz and Sidner's claims about the effect of discourse structure on the search for anaphoric antecedents (Moser and Moore do not propose modifications to Grosz and Sidner's theory in this respect) we have to specify how RDA structure drives focus stack construction. In Grosz and Sidner's theory, the pushing and popping of focus spaces is driven by the intentional structure: a new focus space is pushed on the stack whenever the discourse introduces a new Discourse Segment Purpose (DSP) subordinate to the present one, and the current focus space is popped when the associated DSP is satisfied. Because RDA analyses of a discourse into segments are also based on a notion of intention derived from Grosz and Sidner, a very simple (in fact, possibly *too* simple) way of specifying focus space update in terms of RDA structure can be derived from the following partial mapping between RDA notions and G&S's intentional structure proposed by Moser and Moore :

1. We only have a DSP when we encounter an intentional substructure: i.e., every DSP must be associated with a core.
2. Constituents of the RDA structure that do not include cores - i.e., clusters (see above) - do not introduce DSPs.

The first principle means that a new focus space should only be pushed on the stack when a core is recognized; i.e., only RDA-segments (discourse spans expressing an intentional relation with a core and one or more contributors) are also segments in the G&S sense. The second principle states that discourse spans only connected by informational relations (clusters) do not affect the attentional state.

The segment structure that we would derive from the RDA analysis in Figure 1 by following these principles would be as in Figure 2. Because informational relations by themselves don't give rise to intentional segments, the informational clusters 1.1-1.2 and 3.1-3.2 are not identified as segments in the sense of Grosz and Sidner: i.e., no new focus space is pushed on the stack in this case. In particular, the unembedded core in 1.1-1.2 is not treated as a separate focus space even if it constitutes a cluster, since it expresses (part of) the DSP associated with the overall RDA-segment.

The claim that only intentional relations result in new focus spaces being pushed on the stack already results in different predictions concerning the search for anaphoric antecedents than analyses similarly based on RST-like repertoires of relations, but in which all subordinating RST relations are taken to limit accessibility. This is therefore the first interesting claim to test.

### 3.2 Embedded Segments and The Relation between G&S's Theory and Other Theories of Anaphoric Accessibility

An RDA-style analysis assigns to a discourse a much more detailed structure than the one we would expect to see on the basis of Grosz and Sidner' ideas. In RDA, each clause is treated as a distinct discourse unit, whereas in a G&S-style analysis, multiple sentences are often chunked together without any specific relations between them. Furthermore, G&S make no distinction between cores and contributors, and only allow two intentional relations, whereas in RDA many types of intentional relations are possible. This wealth of information is both a problem and an opportunity. A problem in that even if we only consider RDA segments as candidates for segments in the G&S sense, we still have a number of possible candidates for DSP status to choose from. (Notice that Moser and Moore's principles leave open the possibility that not all RDA segments are associated with a distinct DSPs, hence that the mapping from G&S-segments to RDA-segments is not 1:1.) An opportunity in that we can compare Grosz and Sidner's claims with claims made within the RST framework, as already seen above.

The issue we focused on in this study are embedded segments. Moser and Moore themselves raise the question of how to treat embedded *cores*– spans of text that, while expressing the DSP of an RDA-segment, have in turn the complex structure of an RDA-segment. Moser and Moore do not analyze this problem in detail; but several examples in which the antecedent of a pronoun is contained in an embedded nucleus, and this nucleus expresses an intentional relation, are discussed by Fox (1987) (p. 101), and in practice, we found a number of such cases in the Sherlock corpus. A related issue, not raised before in the RDA literature, is how embedded and not embedded *contributors* should be treated The issue of embedded segments is of interest from the point of view of G&S's theory, both in that it can shed some light on the notion of DSP, and because it raises some questions about whether the attentional state really works as a stack.

The issue of embedded segments is also interesting as a way of comparing G&S's claims with Fox's claim that it is entities in 'active' or 'controlling' propositions that are accessible for pronominal reference.[7] This claim can be reformulated in focus stack terms as suggesting that subordinated focus spaces associated with the contributors in an RDA-segment are not immediately popped, but stay on the stack until the RDA segment of which they are a part is completed. (For example, in the case in Figure 1 and 2, should segment 2.1-2.2 be popped as soon as we are done processing it, or should it remain on the stack until the whole RDA-segment is over, given that it participates in the intentional relation that determines the superordinate segment?)

Embedded cores are also a good topic of comparison between G&S's theory and Veins Theory. Although VT is formulated in RST terms, it can be interpreted as stating that the antecedents introduced in RDA segments that are themselves cores of superordinate segments remain on the stack even after the RDA-segment of which they are a part is completed. So, for example, the antecedents associated with the main RDA segment in Figure 1–those introduced in discourse units 1.1, 1.2, 3.1, and 3.2–would remain on the stack if this segment was in turn the core of an embedding segment.

### 3.3 Stacks Versus Caches

A further reason why the issue of embedded segments is interesting is that while Fox's hypothesis is not entirely incompatible with G&S's views (embedded contributors could be seen as being related by a **satisfaction-precedes** relation) it does stretch the sense in which the attentional state can be

[7] A proposition is ACTIVE if it's part of the same RST scheme as the proposition in which the pronoun occurs; whereas a proposition is CONTROLLING if it's part of a scheme which dominates the scheme in which the pronoun occurs.

interpreted as a ‘stack’. In a discourse with the structure in Figure 1, for example, in order for the contributor 2.1-2.2 to stay on the stack while processing 3.1-3.2, while at the same time ensuring that the material in 3.1-3.2 updates the appropriate focus space (the one which was introduced to store the material in 1.1-1.2–see Figure 2), we need to assume fairly complex sequences of stack operations. More precisely, we need to hypothesize that when segment 2.1-2.2 is completed, the order on the stack of the focus spaces for 1.1-1.2 and 2.1-2.2 is temporarily reversed, which requires two pops and two pushes. While G&S themselves assume auxiliary stacks and the like, it is clear that the more complex the operations, the less attractive the model.

The obvious question is whether these problems disappear when the stack is replaced with a model of the attentional state like that proposed in (Walker 1996, 1998). Walker claims that the attentional state is best viewed as a *cache* rather than as a stack. Her proposal is motivated by three problems with the stack model. First of all, she argues, the stack model cannot explain why the size of embedded segments appears to affect the accessibility of antecedents on the stack. She exemplifies this point by means of the contrast between (2) and (3) . ((Walker 1996), p. 256, Dialogues A and B.) In (2) (which is part of the transcript of a call to a radio show about financial advice (Pollack, Hirschberg, and Webber, 1982), the interruption in b.-d. by the radio show host (H) doesn’t seem to make the previous segment inaccessible: caller C can refer in e. to an entity (the daughter) introduced just before the interruption with a pronoun. However, the less acceptable (3) seems to indicate that the felicitousness of such continuations depends on the length of the intervening segment. In this modification of the previous example, where a further question/answer pair has been added, the continuation is much less felicitous, whereas according to the stack model, the attentional state while processing (3g) is identical with the attentional state while processing (2e).

(2) a. C: Ok Harry, I’m have a problem that uh my-with today’s economy my daughter is working,
b. H: I missed your name.
c. C: Hank.
d. H: Go ahead Hank
e. C: as well as her husband
f. They have a child
g. and they bring the child to us every day for babysitting.

(3) a. C: Ok Harry, I’m have a problem that uh my-with today’s economy my daughter is working,
b. H: I missed your name.
c. C: Hank.
d. H: Is that HANK?
e. C: Yes.
f. H: Go ahead Hank
g. C: as well as her husband
h. They have a child
i. and they bring the child to us every day for babysitting.

The second phenomenon that, according to Walker, is not easy to explain with a stack model is the function of Informationally redundant utterances (IRUs). These are utterances that ”... realize propositions already established as mutually believed in the discourse” (Walker 1996, p. 257). According to Walker, the fact that they are reintroduced means that they are not in fact accessible; the function of IRUs, then, is to put them back (on the cache).

Walker's third piece of evidence against the stack model is the fact that CFs are often 'carried over' segment boundaries: this suggests that the antecedents introduced in a segment do not immediately disappear when that segment is concluded, as we would expect if the corresponding focus space were popped, but stay on until they have been replaced, as it happens in a cache.

In (Walker 1996), it's not clear what should be the contents of cache elements; in (Walker 1998), however, it is suggested that the cache should contain the $n$ discourse entities which has been mentioned more frequently recently.

# 4 Methods

## 4.1 Evaluation Metrics

As said above, the reason why models of discourse structure have been studied in connection with anaphora resolution (and generation) is that they claim to restrict the search for anaphoric antecedents. The 'goodness' of a particular model depends therefore on two measures:

- accessibility: whether the antecedent of an anaphoric expression is on the stack at the moment the anaphoric expression is encountered;
- ambiguity: how many distractors are on the stack at the moment that expression is encountered - i.e., how restrictive the focus stack update mechanism is.

These is an obvious tension between the two measures: we can make all antecedents accessible by leaving them all on the stack, and we can make an anaphoric expression completely unambiguous by not keeping anything in the stack. The 'best' model will be therefore the one with the best trade-off between these two measures. In order to evaluate the extent to which the data support a theory of the attentional state like the focus stack theory, we need to compute both of these measures.

The accessibility of an anaphoric antecedent is simple to measure. Our measure of the ambiguity of an anaphoric expression $a$ is more complex, and depends on both the number of entities on the stack that match it, and the (inverse of) the distance between these entities and the anaphoric expression (a matching antecedent further away will be less of a distractor, since according to Grosz and Sidner, a closer antecedent will be preferred). We call this measure PERPLEXITY:

$$Perplexity(a) = \sum^{n} m(n,a)(\frac{1}{distance(n,a)})$$

where $n$ is the number of elements on the stack; $m(n, a)$ = 1 if discourse entity $n$ matches $a$ (see below), and 0 otherwise; and $distance(n, a)$ = 1 if discourse entity is in the same focus space as $a$, 2 if in the previous focus space, etc.

## 4.2 The Corpus

What we call the Sherlock corpus is a collection of tutorial dialogues between a student and a tutor, collected within the Sherlock project (Lesgold et al. 1992). The corpus includes seventeen dialogues between individual students and one of 3 expert human tutors, for a total of 313 turns (about 18 turns per dialogue), and 1333 clauses. The student solves an electronic troubleshooting problem interacting with the Sherlock system; then, Sherlock replays the student's solution step by step, criticising each step. As Sherlock replays each step, the students can ask the human tutors for explanations. Student and tutor communicate in written form.

The Sherlock corpus was previously annotated using RDA to study cue phrases generation (Moser and Moore 1996a; Di Eugenio et al. 1997). The research group which proposed RDA discusses the following reliability results (Moser and Moore 1996a). 25% of the corpus was doubly coded, and the $\kappa$ coefficient of agreement was computed on segmentation in a stepwise fashion.[8] First, $\kappa$ was computed on agreement at the highest level of segmentation. After $\kappa$ was computed at level 1, the coders resolved their disagreements, thus determining an agreed upon analysis at level 1. The coders then independently proceed to determine the subsegments at level 2, and so on. The deepest level of segmentation was 5; the $\kappa$ values were .90, .86, .83, 1, and 1 respectively (from level 1 to 5).

The Sherlock corpus was converted into an XML format for the present tests. A portion of the annotation of the example in Figure 1 is shown below.

```
<rda-relation id="jg1-06-04-09-g" type="intentional">
  <core id="jg1-06-04-05-g" type="seGment">
    <rda-relation id="jg1-06-04-05-g-dupl" type="cluster">
      <cue id="jg1-06-04-c" type="temporal">Before</cue>
      <first id="first_jg1-06-04-05-r">
      <action id="jg1-06-04-a">
         troubleshooting inside the test stations</action>
      </first>
      <second id="jg1-06-04-05-r" info-rel="step2-step1">
      <matrix id="jg1-06-05-m">
      it is always best
        <action id="jg1-06-05-a">to eliminate both the UUT and
           TP</action>
      </matrix>
      </second>
    </rda-relation>
  </core>
  <contributor id="jg1-06-04-07-r" inten-rel="convince"
         info-rel="act-reason">
    <rda-relation id="jg1-06-06-07-g" type="intentional">
      .....
    </rda-relation>
  </contributor>
  <cue id="jg1-06-08-c" type="sequence">Also</cue>
  <contributor id="jg1-06-04-09-r" inten-rel="convince"
    info-rel="act-reason">
    <rda-relation id="jg1-06-08-09-g" type="cluster">
      ....
    </rda-relation>
  </contributor>
</rda-relation>
```

Figure 3: A portion of the annotation of the example Fig. 1

### 4.3 Anaphora Annotation

We annotated about half of the Sherlock corpus for anaphoric information, using a much simplified version of the annotation scheme developed by the GNOME project (Poesio 2000b), which is based on the MATE meta-scheme (Poesio et al. 1999). More specifically, we marked each NP in the corpus, specified its NP type (proper name, pronoun, the-np, indefinite NP, etc) and its grammatical features (person, gender, number), and then we marked all 'direct' anaphors between these NPs (i.e.,

[8] It is unknown to us whether $\kappa$ was also computed on clusters, and on the specific informational relations used.

no bridges). This scheme has good results for agreement (Poesio 2000a) and has already been used for studying anaphoric accessibility (Poesio et al. 2000). We annotated a total of 1549 NPs, 507 of which were anaphoric.

One problem we had to address was that in the RDA annotation, only tutor turns had been annotated (because the students' questions are very short), but many of the antecedents of anaphoric references in the corpus were discourse entities introduced in the preceding student turn asking the question. Following Fox, we included the first elements of such adjacency pairs (the student turns) a part of the accessibility space for anaphoric expressions in the tutor turn To do this, we enclosed each student turn in a special `student-turn` element, marked the NPs it contained, and made this turn by associating with the student turn a special focus space which would be on the stack when processing the tutor turn. The antecedents introduced in turns further away–e.g., in 'tied' adjacency pairs (Fox 1987)–are not available; we counted the anaphoric expressions whose antecedent was unaccessible for these reasons and factored them out.

A large proportion of the anaphoric expressions whose antecedent is not on the stack are proper names. Because these expressions can be argued not to access the stack to find their antecedent, we also counted them separately.

### 4.4 Computing Accessibility and Perplexity

We evaluated the different proposals concerning the extraction of available antecedents from an RDA structure by running a script over the annotated corpus that simulates focus space construction according to each hypothesis (we discuss the interpretation of the methods in the following section). When the evaluation script encounters an anaphoric expression, it attempts to find its annotated antecedent in the focus spaces that are on the stack according to the chosen stack update method.

In order to measure perplexity, the script also attempts to find all antecedents that match the anaphoric expression (DISTRACTORS). The search for distractors depends on the type of anaphoric expression and is based on heuristics. If the anaphoric expression is a pronoun, an antecedent matches counts as a distractor if it grammatically agrees with the anaphor. If its a description, the script finds the head noun of both anaphor and antecedent (using heuristics based on the POS tags of the content of the noun phrases), and then attempts to match them (because we are working with a limited domain, it was possible to hand-code the basic lexical relations among nouns). The search for distractors has precision and recall of 88%.

### 4.5 Extracting Focus Space Updates from RDA Structures

We compared various ways of extracting focus spaces from RDA structure by letting the evaluation script take a parameter specifying how RDA structures should map into focus stack operations. The possible values of this parameter (and the corresponding behaviors of the script) are as follows.

1. **All:** Push a new focus space on the stack whenever a non-atomic RDA unit (both intentional segments and informational clusters) is encountered, and pop this focus space when the constituent ends.

   E.g., in Figure 1, push a new focus space for all three constituents of the top segment: 1.1-1.2, 2.1-2.2, and 3.1-3.2.

2. **Intentional Only / Imm Pop**: Only push a new focus space when an intentional segment is encountered; pop it as soon as the segment is completed. This is the simplest mapping from RDA into focus stack operations.

   E.g., in Figure 1, only push a new focus space for segment 2.1-2.2, and pop it as soon as that segment is completed. 1.1, 1.2, 3.1, and 3.2 are just added to the top focus space.

3. **Intentional Only / Delay pop of cores**: Only push a new focus space when an intentional segment is encountered. Pop focus spaces introduced for contributor segments immediately; but keep on the stack the focus space associated with a core sub-segment for as long as its embedding segment stays there. This solution is reminiscent of the idea of 'core percolation' in Veins Theory.
4. **Intentional Only / Partial delay pop of tribs** Treat embedded cores as in the previous version, but in addition, keep focus space introduced for contributors on the stack until the segment in which they occur is completed; i.e., as long as they are 'active' in Fox's sense.

   E.g., in Figure 1, do not pop the focus space for segment 2.1-2.2 before processing 3.1-3.2.

# 5 Results

## 5.1 The Distinction between Intentional and Informational Relations

We look first at the impact of the distinction between intentional and informational relations, and Grosz and Sidner's claim that discourse structure is only affected by intentional information.

The following table shows the impact of the distinction on accessibility, i.e., the percentage of anaphoric antecedents which can be found on the stack in each case. The line indicated as 'All' shows the percentage of antecedents which are accessible when both informational and intentional relations push new focus spaces on the stack: 'OK' indicates the number of anaphoric antecedents which are accessible, 'NO' indicates the number of antecedents which are not accessible, 'Out of AP' the cases in which the antecedent is not accessible because it's outside the current Adjacency Pair, and 'PN' the number of cases in which the antecedent is not accessible but the anaphoric expression is a proper name (and can access its denotation through long term memory rather than the stack). The line 'Intentional Only' indicates the percentages for the case when only RDA-segments result in a focus space being pushed.

| | OK | NO | Out of AP | PN |
|---|---|---|---|---|
| **All**: | 199 | 74 | 63 | 158 |
| **Intentional only**: | 280 | 20 | 63 | 131 |

The table shows that separating intentional segments (that introduce new focus spaces) from informational clusters (that don't) makes more antecedents accessible; the result is highly significant by the $\chi^2$ Test ($\chi^2 = 29.47, p \leq 0.001$).

The result just shown is expected: creating more focus spaces (which will then be popped) makes more entities inaccessible. The question is whether the increased accessibility makes the search more difficult by leaving too many distractors on the stack. This is measured by our second metric, perplexity. In the following table, **Baseline** is the model in which no focus space is ever popped; the other models are as above.

| | Perplexity |
|---|---|
| **Baseline** | 3.25 |
| **All** | 1.22 |
| **Imm pop of emb core and trib** | 1.81 |

Both focus stack models reduce the perplexity with respect to the baseline; and crucially, both the model in which all spans are associated with a focus space, and that in which only intentional segments are, bring the perplexity under 2 –i.e., with any of these models, anaphoric expressions are on average unambiguous.

### 5.2 Different Treatments of Embedded Segments

Let us consider next the treatment of embedded cores suggested by Veins Theory, and the treatment of embedded contributors derived from Fox's hypothesis. The first line in the table below shows the percentage of antecedents which are accessible if we treat both embedded cores and embedded contributors as distinct DSPs / separate focus spaces, which are closed off as soon as they are completed. The second line shows what happens if we keep embedded cores on the stack as long as the segment of which they are constituents remains on the stack, as done in Veins Theory. The third line, finally, the results if we treat embedded contributors within an RDA-segment as remaining on the stack until the segment is closed off–i.e., as long as they are 'active' in Fox's sense. In this table we have ignored both cases in which the antecedent is inaccessible but the anaphoric expression is a proper name, and the 63 cases in which the antecedent is inaccessible because it's not in the same adjacency pair (see discussion above).

| | OK | NO |
|---|---|---|
| **Imm pop of emb core and trib** | 280 | 20 |
| **Delay pop of emb cores** | 287 | 16 |
| **Delay pop of emb trib** | 310 | 8 |

The differences are not so large in this case, but the correlation is still significant ($\chi^2 = 6.09, p \leq 0.05$) and in particular there is a highly significant difference between the simplest method for focus stack update (immediately pop embedded segments) and the method in which the popping of embedded contributors is temporarily delayed, and embedded cores remain on the stack (the 'active' model).

The table below shows the impact of these two changes to the treatment of embedded segments on perplexity.

| | Perplexity |
|---|---|
| **Imm pop of emb core and trib** | 1.81 |
| **Delay pop of emb cores** | 1.82 |
| **Delay pop of emb trib** | 1.89 |

As we can see, neither of these treatments results in a significant difference from the simplest model in which embedded segments are immediately popped.

### 5.3 A Simple Cache Model

We also considered a simple version of the cache model proposed in (Walker 1998). This model simply keeps the latest $n$ discourse entities in the cache; whenever an entity is referred to anaphorically, it takes the place in the cache of the oldest element there.

We tried several cache sizes (6, 10, 12, 18), but got worse balance between accessibility and perplexity than with the stack models previously discussed. With a cache of size 6 we get a very good perplexity of 1.79, but less than half the accessibility of the best focus stack model; whereas with a cache of size 18 we get good accessibility, but with a perplexity of 2.47 (above 2).

## 6 Discussion and Conclusions

We can divide the concerns addressed in this work in two groups: which way of using an RDA annotation to test G&S's theory of the global focus works best, and how well the predictions of the theory are verified. A subtopic of the first is whether we can use ideas from RDA to make G&S's theory–and in particular, the notion of DSP–more specific. We consider these questions in turn.

### 6.1 RDA Analysis and Focus Stack Updates

Grosz and Sidner's paper does not tell us how we can identify the DSPs in a discourse. In this respect, our first interesting result is that we get a significantly better characterization of the attentional state if we only associate DSPs with cores: i.e., if new focus spaces are only pushed on the stack when an intentional relation in the RDA sense is observed, and not if only informational relations are. This result in a sense validates the distinction introduced by Moser and Moore, and is especially interesting when compared with Fox's proposals (Fox 1987). According to Fox, informational relations affect accessibility as well (although Fox's study is not concerned with accessibility but with pronominalization, as discussed shortly). It is also interesting to contrast this result with those obtained with Veins Theory (Ide and Cristea 2000), where we also do not find a distinction between informational and intentional relations (although nuclei are assigned a special role not unlike that of cores here).

### 6.2 Evaluating Grosz and Sidner's Theory

If we use the RDA annotation scheme as a guide to identifying DSPs in the sense just discussed, all focus stack update methods we considered clearly reduce the perplexity of anaphoric expressions, and even with the version which keeps embedded contributors on the stack until the intention associated with the core is achieved, we still find a perplexity lower than 2. And even the simplest way of using RDA segments to drive focus stack update (immediate pop) leads to a version with a reasonable performance from the point of view of accessibility as well: just 20 anaphoric antecedents out of stack, and a perplexity of 1.81.

However, the best compromise between accessibility and perplexity is obtained when embedded cores stay on the stack, and embedded contributors are only popped when an RDA-segment is closed. If we adopt the version of the mapping according to which embedded contributors stay on the stack until the RDA-segment to which they belong is completed, only 8 anaphoric expressions have their antecedent outside the currently open focus spaces. Of these, five are cases of definite descriptions that might be viewed as referring deictically to parts of the circuit, one is a cataphoric discourse deixis, one is a temporal deixis (*at this point*), and one is an expression whose interpretation is very ambiguous. In other words, adopting these rules for opening and closing focus spaces, virtually all anaphoric antecedents are accessible.

The problem is that this method for focus stack update, inspired by Fox's notion of 'active' proposition and by Veins Theory's view of 'core percolation,' can only with difficulty be reconciled with the idea that the focus space is a stack. Leaving these focus spaces open means that the attentional state cannot be properly seen as a stack anymore, but has to be seen as a sort of list, since there is no guarantee that the focus space associated with the core will be the last element on the stack. Figure 4 shows an example in which it is necessary to leave the embedded contributor 24.13-24.14 on the stack, in order to solve *the "other" voltage* to *some other voltage*; at the end of the segment, this focus space has to be removed while leaving the focus space for the core on top of the stack. Furthermore, as said earlier, in the case of Figure 1 we need either series of pops and pushes, or a 'discontinuous focus space' to make the embedded contributor accessible without eliminating the principle that new entities are added to the focus space on top of the stack.

24.13a Since S52 puts a return (0 VDC) on it's outputs
24.13b when they are active,
24.14 the inactive state must be some other voltage.
24.15 So even though you may not know what the ”other” voltage is,
24.16 you can test to ensure that
24.17a the active pins are 0 VDC
24.17b and all the inactive pins are not 0 VDC.

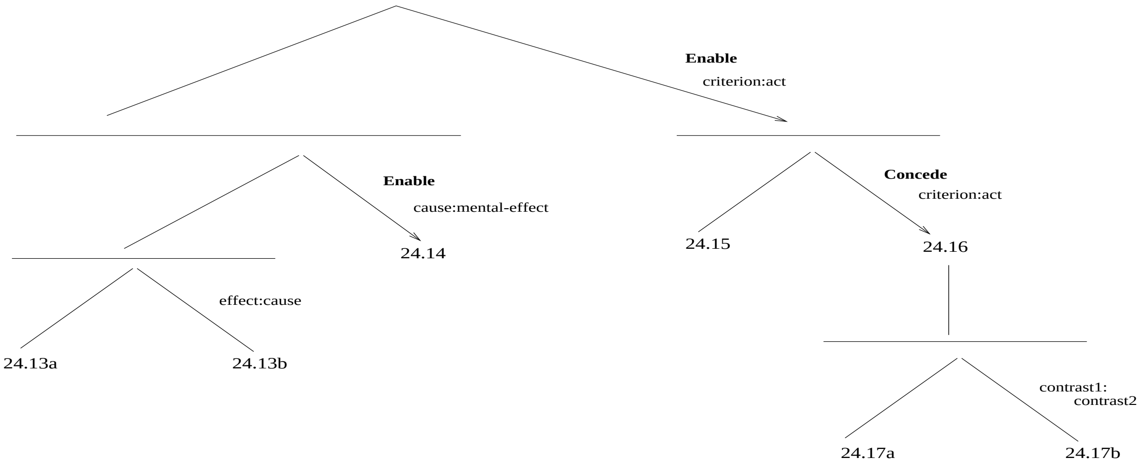

Figure 4: A contributor that precedes the core

### 6.3 A Cache Model?

As said above, the complications above suggest that a cache might be a more natural model of the way accessibility works. However, we also saw that if we replace the focus stack with a cache containing discourse entities, as in the model proposed by (Walker 1998), we get worse results than with the stack models. We are experimenting with models in which the elements of the cache are focus spaces, which is also compatible with the ideas discussed in (Walker 1996).

## 7 Related Work

### 7.1 Fox

Fox (1987), although only concerned with references to singular and human antecedents, is perhaps the most extensive study of the effects of discourse structure on anaphora in both spoken and written discourses. Fox uses different methods for analyzing the two genres: she uses concepts from Conversation Analysis, and in particular the notion of Adjacency Pair, for spoken conversations, and RST to analyze written texts. Her main proposal about written texts is as follows:

> A pronoun is used to refer to a person if there is a previous mention of that person in a proposition that is ACTIVE or CONTROLLING; otherwise a full NP is used.

(Where a proposition is ACTIVE if it's part of the same RST scheme as the proposition in which the pronoun occurs; whereas a proposition is CONTROLLING if it's part of a scheme which dominates the scheme in which the pronoun occurs.)

Fox's proposals concerning pronominalization apply less well to references to objects (and even in her corpus there are many references for which the hypothesis above would licence the use of a pronoun are actually realized by a definite NP, which she explains by arguing that the principle above is only one of many interacting principles that determine the realization of a NP); nevertheless, she makes a lot of compelling points about structure. In particular, she makes it very clear that active propositions should be accessible for as long as the scheme is open; and produces several examples showing that material introduced in active embedded nuclei is accessible. Fox didn't find references inside active embedded satellites (but then again none of these is made via a pronoun in our corpus). In addition, our study suggests that proper names behave differently from definite descriptions in that the former are much less sensitive to discourse structure than the latter, so the two classes should not be conflated like Fox does; and not separating informational relations from intentional ones restricts too much the range of accessible antecedents, even if it may be correct as far as pronominalization is concerned.

### 7.2 Veins Theory

Veins Theory (VT) (Cristea et al. 1998, 2000; Ide and Cristea 2000) is a recently proposed theory of the effect of discourse structure on anaphoric accessibility, which relies on RST for its definition of discourse structure, and whose predictions have been tested using an RST annotated corpus of newspaper texts (Cristea et al. 2000). The propositions accessible to an anaphoric expressions are computed by an algorithm that operates directly over an RST tree and involves two steps: a bottom up step in which the 'heads' of each node in the tree are computed (where the head of a non-terminal node is the concatenation of the heads of its nuclear daughters) followed by a top-down computation of the VEIN EXPRESSIONS. The crucial idea of VT is that material introduced in nuclear nodes 'percolates up' veins, where veins are paths in the tree all of whose arcs connect nuclear nodes; the antecedents introduced in any node along the vein are accessible from all the nodes of the subtree which has the top of the vein to which that node belongs as its root. The second idea of the theory is that antecedents introduced in a satellite node to the left of a nucleus remain accessible to all nodes controlled (in Fox's sense) by that nucleus.

In some respects, the proposal presented here can be viewed as a generalization of the proposals of VT: the material introduced in core constituents percolates up in a similar way, but we also allow antecedents introduced by embedded contributors to the *right* of the nucleus to be accessible as long as these contributors are active (in VT, only binary trees are considered).

The one point of contrast between the two theories is that in our proposal, we do not consider all nuclei, but only the nuclei of intentional relations. We have seen that treating informational relations as introducing focus spaces makes a big difference in terms of accessibility; the difference is significant even if we allow these additional relations to remain open as long as the dominating relation is open. The crucial case distinguishing the two theories are trees with the structure in Fig 5: in our theory X would be available as an antecedent of Y, whereas in Veins Theory it wouldn't.

## References

Carletta, J., A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. H. Anderson: 1997, 'The Reliability of a Dialogue Structure Coding Scheme'. *Computational Linguistics* **23**(1), 13–32.

Cristea, D., N. Ide, D. Marcu, and V. Tablan: 2000, 'Discourse Structure and Co-Reference: An Empirical Study'. In: *Proc. of COLING*. Saarbruecken, pp. 208–214.

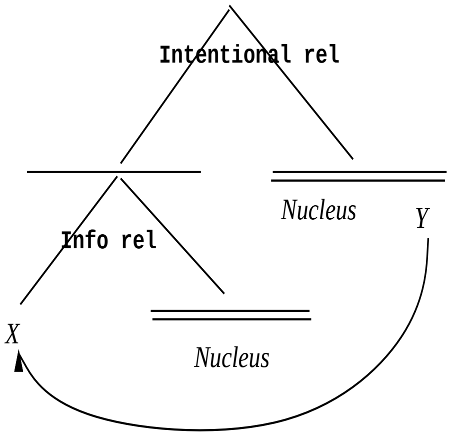

Figure 5: A tree distinguishing Veins Theory from the proposals discussed in this paper

Cristea, D., N. Ide, and L. Romary: 1998, ‘Veins Theory: A Model of Global Discourse Cohesion and Coherence’. In: *Proc. of COLING*. Montreal, pp. 281–285.

Di Eugenio, B., J. D. Moore, and M. Paolucci: 1997, ‘Learning Features that Predict Cue Usage’. In: *Proc. of the 35th ACL*. Madrid.

Fox, B. A.: 1987, *Discourse Structure and Anaphora*. Cambridge, UK: Cambridge University Press.

Grosz, B. J. and C. L. Sidner: 1986, ‘Attention, Intention, and the Structure of Discourse’. *Computational Linguistics* **12**(3), 175–204.

Ide, N. and D. Cristea: 2000, ‘A Hierarchical Account of Referential Accessibility’. In: *Proc. of ACL*. Hong Kong.

Lesgold, A., S. Lajoie, M. Bunzo, and G. Eggan: 1992, ‘SHERLOCK: A coached practice environment for an electronics troubleshooting job’. In: J. Larkin and R. Chabay (eds.): *Computer assisted instruction and intelligent tutoring systems: Shared issues and complementary approaches*. Hillsdale, NJ: Erlbaum, pp. 201–238.

Mann, W. C. and S. A. Thompson: 1988, ‘Rhetorical Structure Theory: Towards a Functional Theory of Text Organization’. *Text* **8**(3), 243–281.

Marcu, D.: 1999, ‘Instructions for Manually Annotating the Discourse Structures of Texts’. Unpublished manuscript, USC/ISI.

Moore, J. and M. Pollack: 1992, ‘A problem for RST: The need for multi-level discourse analysis’. *Computational Linguistics* **18**(4), 537–544.

Moser, M. and J. D. Moore: 1996a, ‘On the Correlation of Cues with Discourse Structure: Results from a Corpus Study’. Unpublished manuscript.

Moser, M. and J. D. Moore: 1996b, ‘Toward a Synthesis of Two Accounts of Discourse Structure’. *Computational Linguistics* **22**(3), 409–419.

Moser, M., J. D. Moore, and E. Glendening: 1996, ‘Instructions for Coding Explanations: Identifying Segments, Relations and Minimal Units’. Technical Report 96-17, University of Pittsburgh, Department of Computer Science.

Nakatani, C. H.: 1996, 'Discourse Structural Constraints on Accent in Narrative'. In: J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg (eds.): *Progress in Speech Synthesis*. New York, NY: Springer Verlag.

Nakatani, C. H., B. J. Grosz, D. D. Ahn, and J. Hirschberg: 1995, 'Instructions for annotating discourses'. Technical Report TR-25-95, Harvard University Center for Research in Computing Technology.

Poesio, M.: 2000a, 'Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results'. In: *Proc. of the 2nd LREC*. Athens, pp. 211–218.

Poesio, M.: 2000b, 'The GNOME Annotation Scheme Manual'. University of Edinburgh, HCRC and Informatics, Scotland, fourth version edition. Available from `http://www.hcrc.ed.ac.uk/ ˜ gnome`.

Poesio, M., F. Bruneseaux, and L. Romary: 1999, 'The MATE meta-scheme for coreference in dialogues in multiple languages'. In: M. Walker (ed.): *Proc. of the ACL Workshop on Standards and Tools for Discourse Tagging*. pp. 65–74.

Poesio, M., H. Cheng, R. Henschel, J. M. Hitzeman, R. Kibble, and R. Stevenson: 2000, 'Specifying the Parameters of Centering Theory: a Corpus-Based Evaluation using Text from Application-Oriented Domains'. In: *Proc. of the 38th ACL*. Hong Kong.

Poesio, M. and Traum, D.: 1997, 'Conversational Actions and Discourse Situations. *Computational Intelligence* **13**(3), 309–347.

Reichman, R.: 1985, *Getting Computers to Talk Like You and Me*. Cambridge, MA: The MIT Press.

Walker, M. A.: 1996, 'Limited Attention and Discourse Structure'. *Computational Linguistics* **22**(2), 255–264.

Walker, M. A.: 1998, 'Centering, anaphora resolution, and discourse structure'. In: M. A. Walker, A. K. Joshi, and E. F. Prince (eds.): *Centering in Discourse*. Oxford University Press, Chapt. 19, pp. 401–435.