Research on Language and Computation **00:** 1–28, 2005.

Discourse Structure and Anaphora in Tutorial Dialogues: an Empirical Analysis of Two Theories of the Global Focus

M. POESIO,[†] A. PATEL[†], AND B. DI EUGENIO[‡]
[†]UNIVERSITY OF ESSEX; [‡]UNIVERSITY OF ILLINOIS AT CHICAGO

Abstract. The recent development of reliable guidelines for discourse structure annotation, and the resulting availability of corpora annotated for discourse structure, have made it possible to subject to rigorous empirical testing the claims of seminal theories about the impact of discourse structure on anaphora. We carried out an empirical investigation of the claims made in two models of the Global Focus–Grosz and Sidner's stack model and Walker's cache model–using a corpus of tutorial dialogues annotated according to Relational Discourse Analysis. We studied how these two models affect both the accessibility of the antecedents and the ambiguity of both pronouns and definite descriptions, examining a variety of stack and cache update strategies and of cache sizes, and paying special attention to the problem of antecedents contained in embedded segments. The best results for the stack model were obtained when DSPs were only associated with intentional relations (i.e., excluding informational relations) and allowing embedded segments to remain on the stack as long as the superordinate segment was open. With the cache model, we found that cache size matters (if the size is less than 10, the model is too restrictive) and that the cache replacement strategy matters even more.

1. INTRODUCTION

Theories of discourse structure have played an important role both in dialogue and NLG research. From a generation point of view, such theories have played an important role in work on text planning, whereas from the point of view of interpretation, one of the main motivations for studying discourse structure is its effect on the search for the antecedents of anaphoric expressions. However, until recently it has not been possible to extensively test such theories either empirically or computationally, because of limitations both in our theoretical understanding of crucial components of such theories (such as the notion of intention) and in the availability of annotated resources.

The recent development of more reliable guidelines for this type of annotation (Nakatani et al., 1995; Moser et al., 1996; Carletta et al., 1997; Marcu, 1999) and the increased availability of corpora annotated for discourse structure (Moser and Moore, 1996b; Carletta et al., 1997; Marcu et al., 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; Cristea et al., 2000; Ide and Cristea, 2000)), reflecting in part the greater availability of corpora annotated in this way, particularly Marcu's Rhetorical Treebank.* Our aim in this work, on the other hand, was to study the claims of two models developed specifically to study the interpretation of anaphoric expressions: the stack model proposed by Grosz and Sidner themselves, and Walker's cache model (Walker, 1998). Both of these models build on the hypothesis made in Grosz and Sidner (1986) that discourse has two levels of 'structure', a global level and a local level, and concentrate on modelling the global level, the so-called GLOBAL FOCUS. We concentrate here on models of the global focus; we reported on our analysis of claims concerning the local focus in (Poesio et al., 2004).

Evaluating Grosz and Sidner's theory used to be a problem, because although it originated a coding manual (Nakatani et al., 1995) 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, which 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. An interesting aspect of these corpora is that they can be used to investigate not just the claims of Grosz and Sidner's theory, but also 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 (Cristea et al., 1998; Cristea et al., 2000; Ide and Cristea, 2000). The Sherlock corpus of tutorial dialogues collected at the University of Pittsburgh (Lesgold et al., 1992) and subsequently annotated according to RDA for different purposes (Moser and Moore, 1996a), is particularly well suited for these purposes. In this study, the Sherlock corpus was used to investigate the claims of Grosz and Sidner's stack model and of Walker's cache model.

The structure of the paper is as follows. First, we quickly review Grosz and Sidner's stack model and Walker's cache model. We then briefly introduce RDA, and discuss how an RDA analysis can be used to analyze Grosz and Sidner's claims about anaphora. Using an RDA analysis raises a number of questions, some of which have been addressed in other work, particularly

* 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.

by Fox and in Veins Theory (Cristea et al., 1998; Cristea et al., 2000; Ide and Cristea, 2000); we briefly introduce these ideas.*

Next, we discuss how we investigated such claims: our evaluation metrics, our corpus, our annotation methods, and how we used the annotated corpus to compute the metrics. It turns out that the RDA annotation can be used in a variety of ways to drive focus stack update; we tested several such methods. We also considered a variety of cache architectures. In the next section of the paper, we discuss the results obtained with these different stack update strategies, and with the several cache architectures we considered.

2. TWO THEORIES OF THE GLOBAL FOCUS

In this section we briefly discuss Grosz and Sidner's and Walker's theories of global discourse structure.

2.1. GROSZ AND SIDNER'S STACK MODEL

According to G&S, discourse has two levels of 'structure', A local level, the LOCAL FOCUS, is updated quickly and often, and is the primary resource for interpreting pronouns and other 'surface' anaphors. Space constraints prevent a discussion of this aspect of the theory, formalized in so-called CENTERING (Grosz et al., 1995).* In addition, discourse has a global structure, the GLOBAL FOCUS, determined by the intentions that the discourse participants intend to convey, or DISCOURSE SEGMENT PURPOSES (DSPs).** In a coherent discourse, these DSPs are all related in an INTENTIONAL STRUCTURE by either **dominance** relations (in case a DSP contributes to the satisfaction of a second DSP) or **satisfaction-precedes** relations (when the satisfaction of a DSP is a precondition for the satisfaction of a second one).

Anaphoric accessibility is modeled by the ATTENTIONAL STRUCTURE of a discourse, which, according to Grosz and Sidner, is a STACK of FOCUS SPACES. G&S propose that 'opening a new discourse segment' really involves pushing on the stack a new focus space, which includes the discourse entities mentioned in that segment. When the segment is completed, its associated focus space is popped, and the discourse entities associated with that focus space are not accessible any more. A further hypothesis is that the pushing and popping of focus spaces on the stack reflects the intentional structure

* It would have been impossible to discuss all relevant studies of discourse structure and anaphora, even just the more recent ones; but unfortunately, space constraints prevent anything more than brief comments in the text about the relation between our work and work on Veins Theory and the work by Tetreault (2005). Other relevant work includes work using Asher and Lascarides' SDRT (Asher and Lascarides, 2004).

* We carried out an empirical analysis of the claims of Centering using methods similar to those used here: see (Poesio et al., 2004).

** 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.

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.

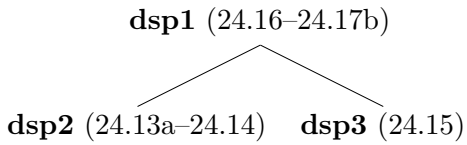

Figure 1. G&S analysis of the structure of a tutorial dialogue fragment.

of a discourse, in the sense that a new focus space is pushed on the stack whenever a new DSP subordinate to the present one is recognized, and a focus space is popped whenever its associated DSP is perceived as having been satisfied.

Consider for example the tutorial dialogue fragment in Figure 1. (This fragment is from the Sherlock corpus of tutorial dialogues used in this study (Lesgold et al., 1992)–the corpus is discussed below.) In this part of the dialogue, the aim of the speaker (a tutor) is to convince the listener (a student) that it makes sense to test that the active pins are 0 VDC. (Note that this aim only becomes explicit in 24.16–24.17a,b.) We take it to be fairly uncontroversial that this goal is the Discourse Segment Purpose for the whole fragment, **dsp1**. It also seems fairly clear that the purpose of utterances 24.13–24.14 and of 24.15 is to provide arguments in favor of this claim, and to discount a possible objection; i.e., that these contributions are made in order to achieve **dsp1**. However, it is not immediately obvious what the intentional structure is, or indeed that there is a single intentional structure; which illustrates what we said earlier about the problems that may arise when attempting a G&S-style analysis of discourse. Depending on one's theory of intentions and formalization of intentional relations, both 24.13–24.14 and 24.15 may be viewed as being part of the same segment as 24.16; alternatively, 24.13–24.14 may be viewed as being related to a separate DSP, **dsp2**; or perhaps both 24.13–24.14 and 24.15 can be viewed as expressing separate DSPs. In Figure 1 we show this last analysis: both 24.13–24.14 and 24.15 express DSPs subordinate to **dsp1**–i.e., each sequence of utterances constitutes a separate discourse segment embedded in the segment associated with **dsp1**. Different intentional analyses will lead to

different predictions concerning the accessibility of *its outputs* in 24.13a (the anchor of the two bridging references *the active pins* in 24.17a and 24.17b) and *some other voltage* in 24.12 (the anchor for *the "other" voltage* in 24.15). The top half of Figure 1 illustrates the segments created as a result of the intentional structure in the bottom half; a separate focus space would be created for each discourse segment and then popped, so this analysis would predict that none of the antecedents of the anaphoric expressions in the fragment (italicized) would be accessible.

Grosz and Sidner's claims about the relation between intentional structure and anaphoric accessibility were illustrated in (Grosz and Sidner, 1986) with a few examples; however, as far as we know, these claims have never been empirically tested in a systematic fashion. Part of the problem is that it is far from obvious how to identify the DSPs in a discourse, as the example just shown illustrates. 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 using an intentional analysis which has been agreed upon by more than one individual.

2.2. WALKER'S CACHE MODEL

The second model of the global focus we tested is the model (Walker, 1996; Walker, 1998). Walker argues that the global focus is best viewed as a *cache* rather than as a stack. Although Walker originally suggested a cache replacement strategy based on intentional structure (Walker, 1996), in her later and more detailed proposal (Walker, 1998) intentions played no role, so this model can also be considered as an alternative to models based on intentions.

Walker's proposal is motivated by a variety of problems with the stack model, of which we'll mention one. 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 (1) and (2) . ((Walker, 1996), p. 256, Dialogues A and B.)

(1)
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.

(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: Is that H A N K?
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.

In (1) (part of the transcript of a call to a radio show reported in (Pollack et al., 1982)), the interruption in b.-d. by the host (H) doesn't seem to make the previous segment inaccessible: caller C can refer in e. with a pronoun to an entity (the daughter) introduced just before the interruption. In (2), by contrast, the addition of a further question/answer pair seems to make the continuation much less felicitous, whereas according to the stack model, the attentional state while processing (2g) should be identical with the attentional state while processing (1e).

24.13a Since [S52]^1 puts [a return (0 VDC)]^2 on [it's outputs]^3
24.13b when [they]$_3$ are active,
24.14 [the inactive state]^4 must be [some other voltage]^5.
24.15 So even though you may not know what [*the "other" voltage*]$_5$ is,
24.16 you can test to ensure that
24.17a [*the active pins*]^6 are 0 VDC
24.17b and all [*the inactive pins*]^7 are not 0 VDC.

Figure 2. State of the cache in the example tutorial dialogue fragment.

In (Walker, 1996), it's not clear what should go in the cache; in (Walker, 1998), however, it is suggested that the cache should contain the n discourse entities which has been mentioned more recently. To illustrate the predictions of the model if the cache is predicted to contain the last n entities, the example text in Figure 1 has been repeated in Figure 2, where each nominal phrase introducing a new discourse entity has been given an integer as a subscript, whereas the subsequent mentions of previously introduced discourse entities are subscripted with the index of that entity. Clearly, in this example every discourse entity is still in the cache when a subsequent mention is uttered, provided that the cache has at least 3 slots.*

3. RELATIONAL DISCOURSE ANALYSIS: A SYNTHESIS OF INTENTIONAL ANALYSIS AND RST

The corpus used in this study was independently annotated according to Relational Discourse Analysis (RDA) (Moore and Pollack, 1992; Moser and Moore, 1996b), a synthesis of ideas from Grosz and Sidner's theory and RST. In RDA discourse structure is determined by intentional structure,

* The assumption here is that only one cache element is used for every discourse entity, no matter how many times that entity is mentioned.

as proposed by Grosz and Sidner: 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.

In RDA, all constituents of a discourse are connected by relations: INFORMATIONAL relations that express a connection between facts and events 'in the world' (such as causal and temporal relations), and / or INTENTIONAL ones that express a discourse intention (such as evidence or concession). While a similar distinction is already present in RST (and SDRT)–SUBJECT-MATTER vs. PRESENTATIONAL relations ((Mann and Thompson, 1988), p. 18)– in RDA the distinction has further significance, in that only spans of discourse tied by intentional relations form proper RDA-SEGMENTs.* Each RDA segment consists of one CORE–the constituent that most directly expresses the speaker's intention** – 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 all subordinating relations have a nucleus and a satellite, whereas in RDA a core and contributors are only identified if a segment purpose has been recognized –i.e., only for RDA-segments.

A feature distinguishing RDA from RST is that in RDA-segments, each contributor is linked to the core by both one intentional relation and one informational relation, whereas in RST, only one relation can obtain between nucleus and satellite. This change was introduced by Moore and Pollack (1992) on the basis of examples like (3).

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

According to Moore and Pollack, the two units can be viewed as being related both by an intentional **evidence** relation (with b as a nucleus, and a as a satellite) and by an informational **volitional cause** one. Furthermore, Moore and Pollack argued that whereas Mann and Thompson claimed that in such cases (which they did observe) one relation had to be chosen, preserving both relations was in fact not only useful to avoid conflicts, but

* A similar distinction is also made in 'structured' versions of Discourse Representation Theory such as SDRT and PTT (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 each other.

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

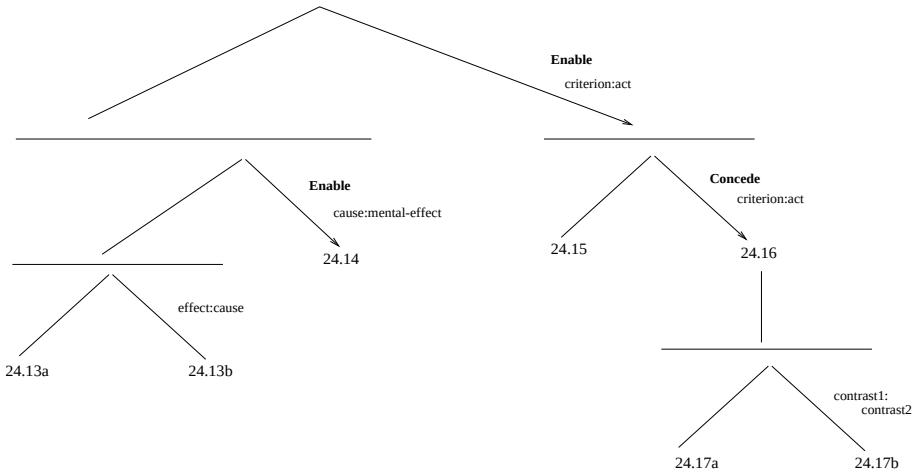

Figure 3. RDA analysis of the example dialogue in Figure 1. In this example both core and contributor are embedded segments, and the contributor precedes the core.

necessary, to account for the flow of inference from both an interpretation and generation point of view. Note that in RDA, in a segment with more than one contributor, contributors may stand in different intentional and / or informational relations to the core.

In RDA, segment constituents may be other (EMBEDDED) segments, 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.

The advantage of RDA over G&S's theory from the point of view of annotation is that RDA is based on a fixed number of relations, like RST. Four intentional relations are identified: **convince**, **enable**, **concede**, and **joint**, together with a larger set of domain-dependent informational relations. In the Sherlock corpus used in this study, 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 3 shows an example RDA analysis, of the small excerpt from the Sherlock corpus of dialogues already discussed in Section 2. The analysis characterizes the text as an RDA-segment whose core spans utterances 24.15-24.17b. This segment has one contributor, spanning 24.13a-24.14. The intentional relation in this case is **enable**.* (Graphically, the core is signaled as the element at the end of the arrow whose origin is the contributor;

* 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).

moreover, the link is marked by two relations, intentional (in bold), and informational.) Both the core and the contributor are further analyzed as embedded RDA segments with a core and a contributor each (we'll return to embedded segments in a moment). The contributor of the first of these additional segments, and the core of the second, are marked as informational clusters. Clusters are marked by one informational relation, but not by intentional relations. **

4. USING AN RDA-ANNOTATED CORPUS TO EVALUATE THEORIES OF DISCOURSE STRUCTURE

Although in (Walker, 1996) cache construction appears driven by intentional structure, the algorithm in (Walker, 1998) only uses information about which entities (forward-looking centers) have been introduced; relations do not matter. Hence, the evaluation of a cache model of this type does not depend on the *relational* information specified by an RDA-style annotation, but only on *entity-level* information–parameters such as the size of the cache or the cache replacement strategy. We discuss these parameters in Section 5.

On the other hand, in order to use an RDA-style annotation 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 to use the RDA structure to determine the DSPs of a discourse–i.e., how RDA structure 'drives' focus stack construction. In this section we discuss, first of all, several ways in which we can use an RDA analysis to identify the DSPs in a discourse–i.e., several ways to decide when a new focus space should be *pushed* on the stack. Secondly, we discuss the problems raised by embedded segments, particularly when they express contributors, and they come before the core. (An example of embedded contributor is found in the RDA analysis of excerpt of dialogue in Figure 1, whose annotation in our corpus is shown in Figure 3.) As we will see, embedded segments raise questions concerning when focus spaces should be *popped*.

4.1. RDA STRUCTURE AND INTENTIONAL STRUCTURE

As said above, in Grosz and Sidner's theory the pushing and popping of focus spaces is driven by the intentional structure of a discourse: a new focus space is pushed on the stack for every Discourse Segment Purpose (DSP) subordinate to the present one, and the current focus space is popped when its associated DSP is satisfied. But although RDA was inspired in part by G&S's work, and an RDA analysis of a discourse into segments is also based on intentions, the two types of structure are not identical. For one

** 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.

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. G&S segments for the discourse in Fig. 1 on the basis of the mapping of RDA-segments into G&S segments proposed by Moser and Moore.

thing, a RDA-style analysis assigns to a discourse a much more detailed structure than an analysis based on 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.

A proposal concerning the mapping from RDA into DSPs was made by Moser and Moore (1996b):

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.

Following these principles we would derive from the RDA analysis in Figure 3 the segment structure in Figure 4. This structure consists of a segment for the top intentional relation, with embedded segments both for its contributor (24.13a-24.14) and for its core (24.15-24.17b). Because informational relations are not interpreted as segments in the sense of Grosz and Sidner, no new focus space is pushed on the stack for the informational clusters 24.13a-24.13b and 24.16-24.17b .

The attentive reader will have noticed that the structure in Figure 4 is not the same as the structure in Figure 1. The segmentation shown in Figure 1 is based on a more 'simple-minded' way of extracting an intentional structure from an RDA structure than that proposed by Moser and Moore. A simpler

way of identifying DSPs can be obtained by simply treating every intentional relation of RDA (**convince**, **enable**, **concede** and **joint**) as a specialized instance of a **dominance** or a **satisfaction-precedes** relation between DSPs. This amounts to associating a DSP to each contributor. More precisely, this second hypothesis concerning the relation between RDA structures and DSPs would work as follows:

1. Associate a DSP with the top segment;
2. Then, associate a subordinate DSP with every contributor, and proceed recursively with the embedded segments.

Notice that this mapping automatically achieves one of the goals of Veins Theory (see below), 'percolating' cores: in the intentional structure resulting from the RDA structure in Figure 3 according to this mapping, shown in Figure 1, the core is not embedded.

Both of these mappings are based on the central hypothesis in RDA: that it is only intentional relations that limit accessibility. This hypothesis clearly leads to different predictions concerning the search for anaphoric antecedents than in a simple-minded RST analysis, in which all subordinating relations (whether informational or intentional) limit accessibility. In the G&S framework, these different views of the mapping become different hypotheses about the decision to push a new focus space on the stack. So the first goal of our computational simulations of focus-stack construction was to compare three types of pushing strategies: push a new focus space (i) for every constituent, (ii) only for contributors, (iii) only for RDA-segments.

4.2. EMBEDDED SEGMENTS: SOME IDEAS FROM RST-BASED THEORIES

The most difficult issue concerning the mapping of RDA-style annotations into focus stack operations is the treatment of embedded segments: cores or contributors which in turn express intentional relations, such as the segment composed of utterances 24.13a-24.14 in Figure 3. As the example in Figure 1 and its analysis in Figure 4 show, there are examples of reference 'inside' an embedded contributor; these cases of anaphoric accessibility are a direct violation of the so-called RIGHT FRONTIER CONSTRAINT (Webber, 1991; Asher and Lascarides, 2004), according to which only entities 'in the right frontier' of the discourse structure tree should be accessible. The problem raised by discourses with this structure is that under either of the ways of using RDA structure to identify DSPs discussed earlier (Figure 1 and Figure 4), segment 24.13a–24.14 would be popped from the stack before the anaphors in 24.15–24.17b– *the "other" voltage*, *the active pins* and *the inactive pins*– are processed (having the antecedents in the global focus is especially important for these last two, as presumably they wouldn't be in the local focus anymore). Moser and Moore already raised 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, but

did not analyze the question in detail. However, to our knowledge, embedded *contributors* have not yet been discussed. Studying accessibility within RDA-style embedded segments is of interest both to shed some light on aspects of G&S's theory such as the notion of DSP, and whether the attentional state really works as a stack; and also in order to compare the predictions of G&S's theory of the attentional state with those of theories formulated in terms of RST notions, such as Fox's or Veins Theory.

Fox's study (Fox, 1987), although only concerned with references to singular and human antecedents, is still perhaps the most extensive study of the effects of discourse structure on anaphora in both spoken and written discourses. Fox uses different theories for analyzing discourse structure in the two genres: concepts from Conversation Analysis (Levinson, 1983), and in particular the notion of Adjacency Pair, are used for spoken conversations, and Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) to analyze written texts. This immediately presented a problem in that tutorial dialogues such as those we studied are a mixture of genres: there is a bit of dialogue, in that the student ask questions that are then answered by the tutor, but most of the content is actually contained in the tutor's answer, which is in effect a monologue. For this reason, we mainly looked at Fox's proposals concerning written texts, cast in terms of RST 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.

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

Fox makes it very clear that active propositions in a rhetorical scheme 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), but she does discuss several examples in which the antecedent of a pronoun is contained in an embedded nucleus, and this nucleus expresses an intentional relation; one of her examples is (4) (Fox 1987, p. 101), in which, according to Fox, the first two utterances (introducing the antecedent *MacPike*) constitute a **circumstance** rhetorical relation which serves as the (embedded) nucleus of an **elaboration** relation with the third utterance (containing the pronoun *she* referring to MacPike) as a satellite.

(4) a. *MacPike* joined the Cal State faculty in 1978 as a lecturer
b. after teaching three years at the University of Hawaii.
c. She received an appointment as an associate professor in 1981. (*The Sun*, July 1983.)

Fox's hypothesis is about the form of reference that should be used rather than whether an entity is accessible or not. Nevertheless, claiming that a pronoun should be used for entities last mentioned in active propositions obviously implies that it is possible to refer to such entities. Her hypothesis is therefore clearly relevant to our concerns, as the contributor in the RDA structure in Figure 3 is active in Fox's sense when the core is being processed. We can in other words interpret her hypothesis as an argument for keeping the contributor on the stack, i.e., not popping it, as long as it's active.

Other ideas relevant to these questions can found in Veins Theory (VT). Although formulated in RST terms, VT can be viewed as stating that 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. For example, the antecedents introduced in units 24.16, 24.17a and 24.17b in Figure 3 would remain on the stack even after the segment whose core is expressed by 24.16-24.17b is closed, and would remain on the stack as long as the embedding segment is a core.

In the case of embedded segments expressing the *core* of a relation, proposing that the associated focus space stays on the stack until the relation is closed is not necessarily inconsistent with Grosz and Sidner's theory; it depends on how we choose to identify the DSPs. For example, under what we called the 'simple-minded' use of RDA structures to identify DSPs, the core of an embedded segment **rs1** may still express the DSP of the embedding segment **rs2**, provided that **rs1** is the core of **rs2**. Thus, in Figure 3, it could be argued that the core of the embedded core, 24.16–24.17b, expresses the overall DSP **dsp1** of this excerpt of dialogue, whereas both the segment covering 24.13a-24.14 and the immediate contributor 24.15 express subordinate DSPs; in fact, this is the structure shown in Figure 1.

However, Fox's proposal is more general: suggesting that antecedents in active units remain accessible is like saying that embedded *contributors* stay on the stack as well. Evidence for this hypothesis are cases like the one in Figure 3, in which *its outputs* in 24.13a appears to be still accessible when 24.17a is processed. Allowing the contributor segment 24.13a-24.14 in Figure 3 and 1 to remain on the stack until the entire RDA-segment of which is part is completed, rather than immediately after 24.14, may not be entirely unmotivated from a Grosz and Sidner point of view. One could use the difference between **dominance** and **satisfaction-precedes**, argue that **enable** is a type of the latter, and then come up with some story about how **satisfaction-precedes** delays stack popping (as far as we are aware, Grosz and Sidner did not specify how this relation affects the stack). However, this type of story may well be a more radical departure from the basic stack mechanism, as it may involve popping intermediate focus spaces, and doing this requires auxiliary stacks. And although Grosz and Sidner already foresee the necessity of an auxiliary stacks mechanism, this in fact means a much

more general model. (Once we allow for a second stack we get the computing power of a Turing machine.)*

These considerations led us to identify four strategies for using the RDA annotation for guiding stack *popping* that we could evaluate:

1. Pop as soon as an embedded segment is closed, the 'strict' interpretation of the correspondence between RDA segments and G&S-style segments;
2. Delay popping for embedded *cores* until the segment is closed (i.e., taking such cores as contributing to the DSP of the embedding segment–an implementation of Veins Theory's proposal);
3. Delay popping for embedded contributors as well (a test of Fox's more general claim making *all* controlling and active propositions accessible;
4. Never pop anything–allow access to all antecedents introduced in the current and previous turn. (A baseline.)

5. METHODS

The effect on anaphoric accessibility and ambiguity of the two theories of the global focus in consideration were analyzed using a corpus of tutorial dialogues previously annotated according to RDA, and in which anaphoric information was then annotated by us. The annotated information was used by Perl scripts (derived from those developed for (Poesio et al., 2004)) which, depending on their input parameters, simulate the construction either of a focus space stack (according to a variety of strategies for pushing and popping elements on the stack on the basis of the annotation) or of a cache (we experimented with caches of different sizes and with different cache replacement strategies). When an anaphoric expression is encountered, separate procedures are used to find all matching antecedents depending on whether the anaphoric expression is a pronoun or a definite description. The scripts then check whether the annotated antecedent is accessible, and whether the particular model of the attentional state being tested makes competing antecedents available (a measure of the degree to which the model restricts ambiguity). At the end, the scripts compute global metrics of accessibility and ambiguity, that we used to evaluate the models we tested. We discuss in turn the data we used, the annotation, the evaluation metrics, and the programs used to compute them.

5.1. DATA

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 consists of 17 dialogues between individual students and one of 3 expert human tutors, for a total of 313 turns (about 18 turns per

* **satisfaction-precedes** is easier to formalize, and has been, in approaches in which intentional structure directly affects accessibility without recourse to a stack, such as SDRT (Asher and Lascarides, 2004) and PTT (Poesio and Traum, 1997).

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. Student queries are very short, whereas tutors' explanations are often very complex. An interesting aspect of this corpus is that it has a mixed nature: in part unstructured dialogue, in part fairly structured explanations by the tutor.

Rhetorical annotation 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 κ coefficient of agreement (Siegel and Castellan, 1988) was computed on segmentation in a stepwise fashion. First, κ was computed on agreement at the highest level of segmentation. After κ was computed, the coders resolved their disagreements, thus determining an agreed upon analysis at level 1, then independently proceeded to determine the subsegments at level 2, and so on. The deepest level of segmentation was 5; the κ 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 study. Unfortunately space constraints prevent showing an example of the annotation.

Anaphoric Annotation We annotated about half of the Sherlock corpus for anaphoric information, using a much simplified version of the annotation scheme developed in the GNOME project (Poesio, 2004). More specifically, we marked each NP in the corpus, annotated its NP type (proper name, pronoun, the-np, indefinite NP, etc) and its agreement features (person, gender, number), and then we marked all 'direct' anaphors between these NPs (i.e., no bridges). This scheme has good results for agreement (Poesio, 2004) and has already been used for studying anaphoric references to the local focus (Poesio et al., 2004). We annotated a total of 1549 NPs, 507 of which were anaphoric. These included 59 pronouns and 137 definite descriptions (*the*- and *that*-NPs and possessives), as well as 227 proper names and 12 other NP types. However, of the 59 pronouns, 8 expressed discourse deixis; we ignored them. Also, because proper names can be argued not to access the stack to find their antecedent, we did not consider them here.

Adjacency Pairs The RDA annotation we are using presents one problem from the point of view of evaluating theories based on intentional structure: only tutor turns had been annotated–since they are the ones that contain examples of explanation–whereas many anaphoric expressions have as antecedents discourse entities introduced either in the preceding student turn asking the question, or sometimes even in turns further back (in the case of so-called 'tied' adjacency pairs (Fox, 1987)). Unfortunately, it is not possible

to recover from the annotation intentional relations between the intentions expressed in a tutor's turn and the student's intentions, or previous DSPs. We therefore made the simplifying assumption that the tutor's turn is dominated by the DSP of the student's turn (viewed as the first part of a question-answer adjacency pair) and made the material introduced in these student turns accessible when processing anaphoric expressions in the tutor turns. Each student turn was enclosed in a special <`student-turn`> element, the NPs it contained were also annotated, and a special focus space was associated with it which was put on the stack when processing the tutor turn. Unfortunately, the antecedents of 38 definite descriptions are introduced in turns further away than the previous student turn–e.g., in 'tied' adjacency pairs (Fox, 1987)–and therefore are not available, which does affect evaluation, as only the stack-based models suffer from this problem. It is important to keep in mind this when looking at the results of the accessibility evaluation.

5.2. EVALUATION METRICS

As said above, the reason for the attention paid to models of discourse structure in research on anaphora resolution and generation is that such models are claimed 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 in the global focus (on the stack or in the cache) when the anaphoric expression is encountered;
- AMBIGUITY: how many distractors are accessible when the expression is encountered - i.e., how restrictive the attentional mechanism is.

These is obviously a tension between these two measures similar to that between precision and recall in NLP research. It is easy to make all antecedents accessible by leaving them all on the stack / in the cache, or to make an anaphoric expression completely unambiguous by not keeping anything in the global attentional state. The 'best' model, however, will be the one with the best trade-off between these two measures.

As the antecedent of an anaphoric expression a either is in the attentional state or isn't, **accessibility**(a) is a binary function with two values, 1 or 0. One of the measures we will use to evaluate model M (cache or stack) with strategy (pushing / popping, cache replacement, etc) S will be $\mathbf{ACC}_S^M$, the percentage of anaphoric antecedents accessible according to that model using that strategy. The computation of an anaphoric expression's ambiguity, however, depends on the model, and its value is not always obvious. For one thing we should distinguish between two measures of ambiguity: counting all discourse entities matching a,[⋆] (**matching**(a)) or only the DISTRACTORS (**distractors**(a)) i.e., the matching elements other

[⋆] Different matching function must be used depending on the type of anaphoric expression; see below.

than the antecedent.** We will use **matching**(a) as our primary measure, but report **distractors**(a) as well. In addition, computing **matching**(a) is pretty straightforward in a cache model, where we just need to count all matching elements in the cache. In a stack model, however, ambiguity depends not just on the number of matching entities on the stack, but also on their position, and on which positions are considered to be causing interference. We compute **matching**(a) in these cases as follows. Let us use $de(a)$ to indicate the discourse entity of which a is a mention, and $FFS(de(a))$ to indicate the first focus space on the stack in which $de(a)$ occurs. According to Grosz and Sidner, a matching antecedent in focus spaces 'below' $FFS(de(a))$ on the stack will not count as a distractor, since a closer antecedent will be preferred. For example, assume the stack is as in Figure 5 when a is encountered, and that $FFS(de(a))$ = **fs2**. Then the only matching antecedents that matter for the computation of **matching**(a) are those in **fs1** and **fs2**.

fs1: $de_1, \ldots de_i$
fs2: $de(a), de_{i+1}, \ldots de_j$
fs3: $de_{j+1}, \ldots de_n,$

Figure 5. Example of $FFS(de(a))$

We use **matching**$(a)_S^M$ to indicate the total number of discourse entities matching anaphor a according to a particular model of the global focus M and strategy S, and use this measure to define two measures of ambiguity: the average **AmbAve** of **matching**$(a)_S^M$ over all anaphors in set A.

$$\mathbf{AmbAve}_S^M = \frac{1}{|A|} \times \sum_{a \in A} \mathbf{matching}(a)_S^M$$

and the percentage **AmbPerc** of anaphors with **matching**$(a) > 1$.

5.3. A CORPUS-DRIVEN SIMULATION OF GLOBAL FOCUS UPDATE AND ANAPHORA RESOLUTION

As said above, the methodology adopted in this study is similar to that used in (Poesio et al., 2004) and (Poesio and Di Eugenio, 2001): we developed Perl scripts that used the annotation to simulate the construction of both focus space models and cache models of the attentional state, and that when encountering an anaphoric expression, would check whether its antecedent

** Notice that **distractors**(a) is not always **matching**(a) - 1, as the antecedent is not always accessible.

was accessible and measured its ambiguity. These scripts took a number of parameters controlling, in the case of the focus space model, how to use the annotation to drive pushing and popping; in the case of the cache model, the size of the cache, and the cache replacement strategy. Separate routines to identify all matching antecedents were developed for pronouns and for definite descriptions. We discuss each of these aspects of the scripts in turn.

Simulating a Stack Model of the Global Focus Using an RDA Annotation

Simplifying a bit, the script reads in the annotated corpus and identifies RDA units (segments and clusters), recording whether they occur as core or contributor of superordinate relations. The values of the input `pushing` and `popping` parameters then determine whether a new focus space is pushed, and when it is popped (see below). Then, whenever an NP is encountered, a new discourse entity is added to the focus space currently on the stack; anaphoric links create equivalence classes of nominal expressions (coreference chains) all realizing the same discourse entity.

The `pushing` parameter can take one of the values `all` (push both intentional segments and clusters on the stack), `simpleminded` (push the top segment and then all contributors), or `intentional` (follow the Moser / Moore proposal and only push intentional segments). E.g., if `all` was chosen as the value of `pushing`, all units identified with a line in the RDA analysis in Figure 3 would be assigned a DSP, which would result in 5 focus spaces being pushed on top of the stack while processing this fragment: one for the contributor of the top intentional relation (covering utterances 24.13a-24.14); one for the embedded contributor, 24.13a–24.13b (even if it's not an RDA segment); then one for the core of the top intentional relation (24.15–24.17b), one for the contributor of this relation (24.15), and one for its core (24.16–24.17b). If `simpleminded` is chosen, three focus spaces would be pushed: two for the topmost constributor as just discussed, and one for the topmost core. Finally, if `intentional` is used, only two focus spaces will be pushed on the stack, one each for contributor and core of the top intentional relation.

The `popping` parameter can take one of the following values:

Pop immediately pop the focus space associated with an RDA segment as soon as it is processed. E.g., with `pushing=all`, when the structure in Figure 3 is processed, the focus space for 24.13a–24.13b would be popposed as soon as 24.13b is processed (thus making all its entities inaccessible); then the focus space for the contributor 24.13a–24.14 as soon as 24.14 is processed, etc.

Delay pop of cores Pop the focus spaces associated to contributors (if any) immediately, but keep on the stack the those associated with core

segments until their embedding segment is completely processed. E.g., in Figure 3, keep on the stack the focus space associated with 24.16–24.17b when the embedding segment 24.15-24.17b is completed (popping only 24.15) and then again when the entire segment 24.13a–24.17b is completed. However, the focus space associated with the contributor segment 24.13a–24.14 is popped as soon as 24.14 has been processed. This strategy is reminiscent of the idea of 'core percolation' in VT.

Partially delayed pop of contributors In addition to keeping on the stack the focus spaces associated with cores, also keep those associated with contributors as long as the embedding segment is active. This strategy implements the more general form of Fox's hypothesis by making antecedents in *all* active propositions accessible, not only those in core segments. This strategy would make the antecedents introduced in 24.13a-24.14 in Figure 3 accessible while processing 24.16–24.17b, so that *the "other" voltage* can be interpreted with reference to *some other voltage*. However, an auxiliary stack would be required to keep around the focus space associated with 24.15–24.17b while removing the focus space associated with 24.13a–24.14 if only material introduced in the core is to be percolated up.

Never As a baseline, we also tested never removing a focus space from the stack once it gets there–i.e., of processing anaphoric expressions without discourse structure-induced restrictions on accessibility.

We obtain 12 possible configurations in total, to which we will refer using the abbreviations in Table I while presenting the results.

Table I. The twelve pushing/popping strategies

Pushing Strategy	**Popping Strategy**			
	Immediate	**Delay Core**	**Delay Contrib**	**Never**
All	A-I	A-DC	A-DT	A-N
Simpleminded	S-I	S-DC	S-DT	S-N
Intentional	I-I	I-DC	I-DT	I-N

Simulating a Cache Model of the Global Focus Using an RDA Annotation

As said above, the version of the cache model in (Walker, 1998) does not depend on intentional structure, so our implementation of the cache update policy does not depend on the RDA annotation; the cache gets updated every time the script encounters a new mention of a discourse entity. The scripts simulating the cache update are affected by two parameters:

Size: The dimension of the cache (any integer). Walker suggests 7 items; we also tested sizes 12, 20 and 25.

Cache Replacement Policy: which discourse entity in the cache should make room (serve as VICTIM) when a new discourse entity has to be added to the cache and all elements of the cache are already occupied. A variety of cache replacement policies have been proposed in Computer Science; the two we tested are:

- **Least Recently Used** (LRU): replace the entity which has been in the cache the longest (this is the policy suggested by Walker);
- **Least Frequently Used** (LFU): replace the entity which has been accessed (i.e., mentioned) the least. In case of a tie, we chose the oldest discourse entity as victim.

Finding the Antecedents that Match

In order to compute **matching**(a) it is necessary to find all discourse entities in the attentional state that match a: the actual antecedent, if it's there, as well as the distractors. In order to do this, we developed two distinct algorithms, one for finding the antecedents matching a pronoun, the other for finding the antecedents which match a definite description. Both algorithms use the same search procedure:

1. With the cache model, all discourse entities in the cache are considered;
2. With the stack model, all discourse entities are considered which are realized in the focus spaces up to and including $FFS(de(a))$

Pronouns The algorithm activated when the anaphoric expression is a pronoun returns all discourse entities in the attentional state realized by a noun phrase that matches the agreement features of the pronoun. Our algorithm has perfect recall: i.e., all antecedents in the attentional state are retrieved. (Of course the algorithm cannot find antecedents if they are not in the attentional state.) Obviously precision is not defined in this case as the algorithm does not attempt to make a choice.

Definite descriptions This procedure is more complicated in that definite description resolution depends in many cases on lexical and commonsense knowledge (e.g., in the domain under considerations, *the signals* may be used to refer to *the inputs*) or at least on being able to interpret acronyms (e.g., *the TS* can be used to refer to *the Test Station*). WordNet is notoriously not very useful in restricted domains (Vieira and Poesio, 2000), but as domain knowledge is very limited in the Sherlock corpus, we simply hand-coded synonyms and hypernyms, and the types of named entities (e.g., that *the A1A3A* is a **card**). The matching algorithm for definite descriptions returns all discourse entities in the attentional state realized with a NP that either

1. Have the same head noun as the definite description; or
2. Have a head noun which is a synonym or a hyponym of the head noun of the definite description; or
3. Are proper names, and whose type matches as in 1. or 2.

This algorithm also achieves perfect recall in the sense discussed for pronouns. In addition, our algorithm records all premodifiers of both the definite description and of all potential antecedents and, in case **matching**$(a) > 1$ for definite description a, checks whether the ambiguity could be resolved using the premodifiers. For example, the two potential antecedents for *the high side* are *the low side* and *the high side*; using premodifiers the matching algorithm can determine that the latter is the more likely interpretation.

6. RESULTS

6.1. THE STACK MODEL

The accessibility and ambiguity values for the 51 pronouns in our corpus with each of the 12 pushing / popping strategies are shown in Table II. There is one row for each of the twelve combinations of pushing / popping strategy considered; for each combination, the row shows, first of all, the value of **ACC** (the percentage of pronouns for which the annotated antecedent was found on the stack), then the percentage of pronouns for which it wasn't.* The table then reports **AmbAve** (the average number of matching antecedents for a pronoun) and the average number of distractors.

Table II. Comparison between stack strategies for pronouns

Comb.	ACC	Not Acc	Amb Ave	Distr avge
A-I	90.2%(46)	7.8% (4)	2.78	1.9
A-DC	90.2%(46)	7.8% (4)	2.78	1.9
A-DT	90.2%(46)	7.8% (4)	2.88	2
A-N	90.2%(46)	7.8% (4)	3.13	2.25
S-I	90.2%(46)	7.8% (4)	2.84	1.96
S-DC	90.2%(46)	7.8% (4)	2.84	1.96
S-DT	90.2%(46)	7.8% (4)	2.98	2.09
S-N	90.2%(46)	7.8% (4)	3.20	2.31
I-I	90.2%(46)	7.8% (4)	3.33	2.45
I-DC	90.2%(46)	7.8% (4)	3.33	2.45
I-DT	90.2%(46)	7.8% (4)	3.58	2.7
I-N	90.2%(46)	7.8% (4)	3.68	2.8

Table II clearly shows that different ways of constructing the focus space on the basis of the annotated RDA information do not affect the accessibility of antecedents for pronouns, but only the number of matching elements to

* There is one pronoun whose antecedent is not included in a focus space: a demonstrative pronoun used for a discourse deictic reference (the antecedents of discourse deixis are not stored on the stack).

decide from. (And because the percentage of ambiguous pronouns increases only slightly, the differences in **Fanaph** are minor.) This is yet another confirmation of the hypothesis that the interpretation of pronouns depends primarily on the local focus (i.e., on the antecedents introduced in the same or the previous sentence) more than on the global focus. The results are quite different for the 137 definite descriptions, however, as shown in Table III, which also shows percentages of ambiguous expressions before and after modifier check.

Table III. Comparison between stack strategies for def. descriptions

Comb.	ACC	Not Acc	Amb Ave	Distr avge	Amb Perc	AmbP Wth Pre
A-I	42.3%(58)	57.7%(79)	0.66	0.28	8.8%(12)	8.0%(11)
A-DC	48.9%(67)	51.1%(70)	0.78	0.34	9.5%(13)	8.0%(11)
A-DT	62.0%(85)	38.0%(52)	0.98	0.47	13.1%(18)	9.5%(13)
A-N	72.2%(99)	27.8%(38)	1.09	0.50	13.1%(18)	9.5%(13)
S-I	52.5%(72)	47.5%(65)	0.77	0.30	8.8%(12)	5.1%(7)
S-DC	55.5%(76)	44.5%(61)	0.97	0.35	9.5%(13)	5.8%(8)
S-DT	70.0%(96)	30.0%(41)	1.07	0.50	13.1%(18)	8.8%(12)
S-N	73.7%(101)	26.3%(36)	1.13	0.52	13.9%(19)	9.5%(13)
I-I	64.2%(88)	35.8%(49)	0.93	0.39	12.4%(17)	8.0%(11)
I-DC	67.1%(92)	32.9%(45)	0.97	0.40	12.4%(17)	8.0%(11)
I-DT	72.3%(99)	27.7%(38)	1.08	0.48	13.9%(19)	9.5%(13)
I-N	73.7%(101)	26.3%(36)	1.10	0.49	13.9%(19)	9.5%(13)

Table III shows that unlike with pronouns, in the case of definite descriptions different stack pushing and popping strategies do result in clear differences in accessibility and ambiguity. As already reported in (Poesio and Di Eugenio, 2001), pushing a new focus space on the stack for every relation, whether informational or intentional, greatly reduces accessibility: on average, with the three configurations with `pushing=all` the percentage of accessible antecedents is 51%, and less than 50% unless all active propositions are kept on the stack; whereas with the three configurations with `pushing=intentional`, the average percentage of accessible antecedents is 67.9%; this difference is highly significant by the χ^2 test (Poesio and Di Eugenio, 2001). The choice of popping strategy also affects accessibility, with all types of pushing strategy. Our results show that with the I-DT configuration, which combines limited, 'intentional' pushing with Fox-inspired delayed popping of both cores and contributors, virtually every antecedent that ever goes on the focus stack is made accessible. (We remind the reader that our current scripts only consider antecedents introduced in the current student question-tutor answer adjacency pair. This limitation affects the results. The antecedents of 36 definite descriptions are introduced in student turns further back, and therefore are inaccessible even if no focus space is ever

popped; whereas the antecedents of 2 more definite descriptions introduced in tutor turns are only accessible via the 'never pop' strategy. So in effect the maximum possible value of **ACC** for the popping strategies **I**, **DC** and **DT** is 72.3%.) At the same time, our new results about ambiguity indicate that the increase in accessibility achieved by keeping active propositions on the stack for longer does not increase the number of distractors overmuch: the value of **AmbAve** for definite descriptions is around 1 with all models, and the average number of distractors stays around .5. The **I-DT** configuration, merging Fox's hypothesis about active propositions with the restriction of segmentation to intentional relations achieves the best overall balance.

In our opinion, the fact that **AmbAve** for pronouns is around 3 does not indicate that the models we are considering are not sufficiently restrictive, but that, as already argued in previous literature, the choice of the preferred interpretation for pronouns does not depend on the global focus, but on other preferences. On the other hand, it is assumed that the global focus should eliminate or reduce ambiguity in the case of definite descriptions, so it was interesting to find that more than 10% of definite descriptions in our corpus had more than one matching antecedent. We checked therefore if ambiguity was prevented by premodification in these cases. The results are shown in the last column of Table III. Premodifiers do not help much the more restrictive pushing strategies, but do reduce the number of ambiguous anaphors by almost 50% with the strategies pushing fewer focus spaces.

6.2. THE CACHE MODEL

The results obtained with the versions of the cache model we tested are shown in Tables IV (for pronouns) and V (for definite descriptions). In each table we show the results obtained with two different cache replacement strategies, and with four different cache sizes. For each combination, we indicate the value of **Acc** and **AmbAve**, and the percentage of distractors.

Table IV. Comparison between configurations of the cache model for pronouns

Repl Policy	**Cache Size**	**ACC**	**AmbAve avge**	**Distractors**
LRU	7	39.2% (20)	3.7	3.27
	12	56.8% (29)	6.2	5.74
	20	70.5% (36)	9.6	5.74
	25	83.5% (42)	12	11.3
LFU	7	13.7% (7)	4.09	3.96
	12	27.4% (14)	6.65	6.37
	20	49.0% (25)	9.51	9.10
	25	62.7% (32)	10.86	10.30

Table V. Comparison between configurations of the cache model for definite descriptions

Repl Policy	Cache Size	ACC	AmbAve avge	Distractors
LRU	7	30.6%(42)	0.50	0.17
	12	55.5%(76)	0.80	0.26
	20	68.6%(94)	1.02	0.35
	25	78.8%(108)	1.20	0.48
LFU	7	13.9%(19)	0.46	0.22
	12	29.2%(40)	0.74	0.38
	20	44.5%(61)	0.98	0.38
	25	48.2%(66)	1.08	0.44

As we can see from these Tables, the setting of both parameters of the models are crucial. The choice of replacement policy has a huge impact: with LRU we obtain accessibility results equal or better than those obtained with the stack models, at least for certain sizes of the cache, whereas replacing the least frequently accessed entity (LFU) results in about half that accessibility. However, cache size is also important: a cache with less than 20 slots has worse accessibility than a stack model (with similar average ambiguity), whereas a cache with more than 25 slots has a slightly better accessibility, although with higher average ambiguity. (Average is particularly high in the case of pronouns, but as said above, arguably this is not a problem for a theory of the global focus.)

7. DISCUSSION

It is of course too early to draw definite conclusions, given also the fairly small size of the sample, but this study raised a number of interesting questions to be investigated in greater depth. Among these, we'll discuss the issue of the identification of intentional structure, the way intentional structure drives focus stack construction, and the comparison between focus stack models and cache models.

7.1. RDA AND DSPS

One of the great problems the field of discourse is facing at the moment is that whereas most post-Gricean theories of language interpretation attribute a great importance to intention recognition, very few such theories are detailed enough to enable an analyst to identify unambiguously intentions in a dialogue. Grosz and Sidner's theory suffers from the additional problem that it doesn't simply require to identify intentions (all contributions to a dialogue like the one in Figure 3 presumably are the result of some intention);

it also requires to identify those intentions that are DSPs, i.e., that affect discourse structure (clearly, not all intentions are DSPs). However, Grosz and Sidner's paper does not tell us how we can identify DSPs among all intentions that can be recognized in a dialogue. Unless we find ways of carrying out such empirical investigations, much work in discourse structure based on an intentional analysis will be of limited relevance to current work on anaphora resolution or NLG. One of the central aims of our investigation was to shed some light on this issue.

An RDA-style analysis gives us some help towards identifying discourse structure-affecting intentions, at least in tutorial dialogues of the sort we are studying, by drawing a distinction between two types of contributions to a discourse: those that are related to previous contributions by informational relations (which supposedly do not affect segmentation) and those that are linked by genuinely 'rhetorical'–i.e., speech act-level–relations. This distinction is not made in RST-based analyses of the effect of discourse structure on anaphora, such as Fox's or Veins Theory. According to Fox, informational relations affect accessibility as well, at least in written texts (although Fox's study is not concerned with accessibility but with pronominalization). In Veins Theory (Ide and Cristea, 2000) we do not find a distinction between informational and intentional relations either (although nuclei are assigned a special role not unlike that of cores here). Our first result is that we do 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. This result validates the distinction introduced by Moser and Moore.

However, we also found that this distinction, like the distinction between the popping methods we consider (see below) only matters for definite descriptions, not for pronouns. With pronouns we get exactly the same accessibility results, and very small differences in perplexity, with all strategies for driving stack construction off the RDA annotation. This result might explain Tetreault's result (Tetreault, 2005) that segmentation does not improve the performance of his pronoun resolution algorithm, LRC. (Interestingly, this does not hold with the cache models, where we find that changes in cache size have a big impact on pronoun perplexity.)

However, our attempt at using an RDA structure to 'drive' focus stack construction also suggests that there are a number of theoretical issues to be addressed in doing this. We concentrated here on the issue of embedded segments, finding that suggestions developed by Fox and in Veins Theory did help in getting a better model–but are not easy to incorporate in a stack model (see below).

7.2. STACK MODELS AND EMBEDDED SEGMENTS

The best compromise between accessibility and perplexity was obtained by leaving embedded cores on the stack, and only popping embedded contributors when an RDA-segment is closed. Whether one is willing to accept our conclusions concerning G&S's theory depends of course on whether they think our use of RDA to study the theory convincing; it's clear however that the interpretation we have adopted is very favorable to G&S. However, we already remarked that this method for focus stack update, inspired by Fox's notion of 'active' proposition and by Veins Theory's idea of 'core percolation,' can only with difficulty be reconciled with the idea that the focus space is a stack. Leaving on the stack the focus space associated with the core 24.15–24.17b in Figure 3 while popping 24.13a-24.14 can only be done with the help of an auxiliary stack, which means a radical increase in the power of the formalism.

7.3. STACK VS CACHE

Another result of our study is that the 'best' cache configuration (**LRU** with 25 places) gives better accessibility than the the 'best' stack pushing / popping strategy. On the one hand, we don't think one should read too much in this, as the cache models were given an unfair advantage in that they were able to access discourse entities introduced arbitrarily far back (within the limits of the cache), whereas because of the limitations of our treatment of accessibility in dialogue, the stack models could only access material introduced in the current turn and the previous turn, because the annotation did not include intentional relations 'across turns': the main difference between the two models are those 38 antecedents of definite descriptions that are found neither in the current nor the previous turn. So, we don't think this study should be used to settle the theoretical issue of whether caches or stacks are the better models of the global focus; rather, the most reliable theoretical results concern the differences between configurations of each model. On the other hand, it must be said that a cache model like the one we tested is clearly much easier to implement, as there is no need for intention recognition, so it is fair to say that our study didn't provide any evidence that an anaphora resolution system, or sentence planning module, incorporating a stack model would behave better than a system which simply implements a cache model.

As far as the best strategy for a cache model is concerned, we found that both parameters we considered–cache size and cache replacement policy–had a huge impact on the results. Perhaps the most surprising result is how much better the **LRU** cache replacement policy (choose as victim the oldest entity in the cache) works than **LFU** (choose as victim the entity mentioned less often). However, the effect of size is also interesting: with a cache size of

7 (the figure one finds mentioned more often–e.g., in Walker's proposal–no doubt as a result of George Miller's seminal study), accessibility is very low, and only gets comparable to that obtained with stack models with a cache size greater than 20. On the other hand, with these sizes we also get a very high perplexity for pronouns (four times the perplexity obtained with the best stack configurations). As we said earlier, we do not think that it's the job of global focus models to narrow down the choice of pronouns, but this difference is worth studying more carefully.

ACKOWLEDGMENTS

We wish to thank our anonymous reviewers for many helpful suggestions, not all of which unfortunately we could incorporate in the last version of the paper for lack of space. A substantial part of this work, including the creation of the corpus, was supported by the EPSRC project GNOME, GR/L51126/01. Massimo Poesio was supported during parts of this project by an EPSRC Advanced Fellowship. Barbara Di Eugenio was supported in part by NSF grants INT 9996195 and IIS 0133123, in part by NATO grant CRG 9731157.

References

Asher, N. and A. Lascarides: 2004, *The Logic of Conversation*. Cambridge University Press.

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.

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., A. K. Joshi, and S. Weinstein: 1995, 'Centering: A Framework for Modeling the Local Coherence of Discourse'. *Computational Linguistics* **21**(2), 202–225. (The paper originally appeared as an unpublished manuscript in 1986.).

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.

Levinson, S.: 1983, *Pragmatics*. Cambridge University Press.

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.

Marcu, D., M. Romera, and E. Amorrortu: 1999, 'Experiments in Constructing a Corpus of Discourse Trees: Problems, Annotation Choices, Issues'. In: *Workshop on Levels of Representation in Discourse*. pp. 71–78.

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.: 2004, 'The MATE/GNOME Scheme for Anaphoric Annotation, Revisited'. In: *Proc. of SIGDIAL*. Boston.

Poesio, M. and B. Di Eugenio: 2001, 'Discourse Structure and Anaphoric Accessibility'. In: I. Kruijff-Korbayová and M. Steedman (eds.): *Proc. of the ESSLLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics*.

Poesio, M., R. Stevenson, B. Di Eugenio, and J. M. Hitzeman: 2004, 'Centering: A Parametric Theory and its Instantiations'. *Computational Linguistics* **30**(3), 309–363.

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

Pollack, M., J. Hirschberg, and B. Webber: 1982, 'User participation in the reasoning process of expert systems'. In: *Proc. of AAAI*. pp. 358–361.

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

Siegel, S. and N. J. Castellan: 1988, *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill, 2nd edition.

Tetreault, J.: 2005, 'Decomposing Discourse'. In: A. Branco, T. McEnery, and R. Mitkov (eds.): *Anaphora Processing: Linguistic, cognitive and computational modelling*, Current Issues in Linguistic Theory. Amsterdam: J. Benjamins, pp. 73–95.

Vieira, R. and M. Poesio: 2000, 'An empirically-based system for processing definite descriptions'. *Computational Linguistics* **26**(4), 539–593.

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.

Webber, B. L.: 1991, 'Structure and Ostension in the Interpretation of Discourse Deixis'. *Language and Cognitive Processes* **6**(2), 107–135.