A formal semantics for situated conversation *

Julie Hunter
Universitat Pompeu Fabra & IRIT, Université Paul Sabatier

Nicholas Asher
IRIT, CNRS

Alex Lascarides
University of Edinburgh

Abstract While linguists and philosophers have sought to model the various ways in which the meaning of what we say can depend on the nonlinguistic context, this work has by and large focused on how the nonlinguistic context can be exploited to ground or anchor referential or otherwise context-sensitive expressions. In this paper, we focus on examples in which nonlinguistic events contribute entire discourse units that serve as arguments to coherence relations, without the mediation of context-sensitive expressions. We use both naturally occurring and constructed examples to highlight these interactions and to argue that extant coherence-based accounts of discourse should be extended to model them. We also argue that extending coherence-based accounts in this way is a nontrivial task. It forces us to reassess basic notions of the nonlinguistic context and rhetorical relations as well as models of discourse structure, evolution, and interpretation. Our paper addresses the conceptual and technical revisions that these types of interaction demand.

Keywords: situated communication, nonlinguistic context, discourse structure, rhetorical relations

1 Introduction

Consider the following commonplace exchange. Suppose a man, Peter, comes home to find his wife, Anne, looking upset. Peter looks at Anne inquiringly, and she says:

(1) Our daughter was sent to her room.

* We gratefully acknowledge support from ERC grant 269427 and Juan de la Cierva fellowship IJCI-2014-22059. We would also like to thank the participants in our summer school courses on situated discourse at NASSLLI 2016 and ESSLLI 2017 for discussion of many points in this paper, as well as three anonymous *Semantics & Pragmatics* reviewers for extensive and very helpful comments. Finally, we thank the editor, David Beaver, for his detailed feedback, which we feel greatly improved the final paper.

Just after she says this, she nods suggestively over her shoulder, and Peter, taking her cue, notices some scratches on the wall behind her. He immediately infers that the scratches are the outcome of an *event* of their daughter scratching the wall and that this event, call it 'e', provides an explanation of the punishment described in (1). The event e can then support discourse continuations, such as (2):

(2) I was cooking dinner.

Here, the inferred event e affects the temporal interpretation of (2): without e, the discourse (1)+(2) implies that Anne cooking dinner and sending her daughter to her room temporally overlap, but this is not the message Anne is conveying.

Let *Scratches* be the discourse above, involving (1)+e+(2). Now imagine a different discourse, but in the same context. Peter comes home to find Anne looking upset, but she utters (3) rather than (1):

(3) I moved the table into the living room this morning.

She nods suggestively, just as in *Scratches*, but then continues with (4).

(4) I had to buy some new paint.

Here again, the inferred wall-scratching event, e', plays a crucial discursive role: if we ignore e', the discourse (3)+(4) is barely coherent. We'll call this discourse *Table*.

The fact that nonlinguistic events can influence the interpretation of a discourse is old news. Nevertheless, the particular *way* in which e and e' contribute to the content of *Scratches* and *Table*, respectively, has not been a central topic of research. In these discourses, the nonlinguistic events do not provide referents for referential or anaphoric expressions; in fact, their relevance need not be signalled by the presence of any particular linguistic expression. Nor does the linguistic unit describe a concurrent nonlinguistic event, a focus of research on modelling multimodal demonstrations (Stojnic et al. 2013, Forbes et al. 2015). The discursive contributions of the nonlinguistic events are rather like those of the italicized clauses in (5) and (6):

(5) Our daughter was sent to her room. *She scratched up the wall.* I was making dinner.

(6) I moved the table into the living room this morning and *I scratched up the wall*. I had to buy some new paint.

The way in which e and e' respectively affect the content and interpretation of *Scratches* and *Table* will look immediately familiar to any researcher who uses rhetorical structure to represent the content of discourse. Our goal in this paper is

in part to show that by extending work on rhetorical structure, which has thus far focused on linguistically-specified discourse, we can develop a model for discourses like *Scratches* and *Table*, in which nonlinguistic events somehow contribute the contents of entire discourse units; that is, in which they do not simply provide referents for, or otherwise specify the interpretation of, context dependent expressions in the linguistic units. At the same time, we also aim to show that despite the intuitive connection between (5) & *Scratches* and (6) & *Table*, such an extension is far from trivial. Using both naturally occurring and constructed examples, we argue that the extensions required to model *Scratches* and *Table* not only introduce technical complications but also require significant shifts in the way that one should conceive of discourse interpretation.

We begin in Section 2 by more carefully circumscribing the particular type of semantic interactions that we aim to model and by situating our contribution relative to other formal and computational work on the nonlinguistic context. In Section 3, we introduce a corpus of naturally occurring data that we will use to supplement constructed examples like *Scratches* and *Table*. Section 4 explores the conceptual consequences of our proposed extension, and in particular, the way that it requires us to reassess certain notions of the nonlinguistic context and of rhetorical relations. Section 5 tackles two problems: first, it shows how modelling our motivating examples requires modifying constraints on discourse salience and the rules governing the shape of allowable discourse structures; second, it provides a dynamic semantic model theory in which nonlinguistic events become part of an interpreted discourse structure all the while changing the world of evaluation. This results in a different view of the role of semantics in a dynamic, nonlinguistic environment. Section 6 discusses further related work on discourse structure.

2 The context sensitivity of *Scratches* and *Table*

Our aim is to develop a model for a kind of contextual interaction that is on the one hand well-known and complementary to many phenomena studied in formal and computational approaches to meaning, but that on the other hand has not been systematically modelled within these fields. Modelling this interaction will involve looking at two directions of contextual influence: the effects of nonlinguistic events on discourse interpretation and the effects of discourse structure on the typing or conceptualization of nonlinguistic events. We introduce the first direction in Section 2.1 and explain that the mechanism by which e and e' contribute to the content of *Scratches* and *Table*, respectively, lies outside the purview of extant models of context sensitivity. Section 2.2 introduces the second direction, but also explains that a detailed study of how discourse influences conceptualization is beyond the scope of this paper.

2.1 From nonlinguistic events to the interpretation of discourse

It is well-known that nonlinguistic events and entities can influence the interpretation of a discourse. They can serve, for example, as referents for referential or anaphoric expressions. Suppose that instead of *Scratches*, Anne had uttered (7) and pointed at the scratches on the wall while saying *this*:

(7) Our daughter was sent to her room for *this*. *It* happened while I was cooking dinner.

In (7), as in *Scratches*, both the scratching event and the scratches that result from it contribute to the content of the discourse. Also as in *Scratches*, Peter must infer that the scratches that he sees on the wall are the outcome of a scratching event e for which his daughter was somehow responsible. This inference is required for Peter to understand *why* his daughter was sent to her room and *what* happened while Anne was cooking — that is, it is required to compute the references for *this* and *it*.

Despite these similarities, examples such as *Scratches* and *Table* force us to confront challenges that have received little attention in referential semantics. A deictic use of a pronoun can simply be assigned an interpretation by an assignment function, thereby allowing us to put aside the question of what factors actually determine its interpretation and instead address questions such as how to model its contribution *given* a certain interpretation. By contrast, there is no expression in *Scratches* that we can associate with the nonlinguistic event of Anne and Peter's daughter scratching up the wall. Nor is there an expression that triggers a search for a nonlinguistic entity. *Scratches* and *Table* therefore raise several questions: (i) How do e and e' come to be semantically relevant? (ii) What is their semantic contribution to their respective discourses? (iii) Where should these contributions be reflected in the logical forms for *Scratches* and *Table*? (iv) And finally, how do e and e' come to be associated with the semantic content that they convey? Extant work on referential semantics is not designed to answer these questions.

Recent work in computational linguistics and robotics also tackles a different set of questions. There are efforts in distributional semantics to learn how to ground referents for words in the visually salient scene (Baroni 2016), while work in robotics has yielded computational systems that can automatically find extra linguistic referents for referring descriptions (e.g., Kranstedt et al. 2006, Matuszek et al. 2012). The study of multimodal interactions in Human Robot Interaction (HRI) has also led to systems that map speech and visual signals into a unified semantic representation of speaker meaning in order to estimate intentions and beliefs (Perzanowski et al. 2001, Chambers et al. 2005, Foster & Petrick 2014). By and large, this work, like work on referential semantics, is concerned with nonlinguistic events that serve as the denotations of linguistic expressions or linguistic phrases, and it therefore targets

a different dimension of multimodal meaning than that exhibited by *Scratches* and *Table*, in which none of the linguistic phrases denote e or e'.

In this paper, we propose that the semantic mechanisms at work in *Scratches* and *Table* are akin to those that have been studied extensively in work on rhetorical structure, although this body of work has so far focussed almost exclusively on semantic interactions between linguistically-specified contents. In answer to question (i) (*How do e and e′ come to be semantically relevant?*), there is no expression that signals the relevance of e or e'; it is rather the need to find a coherent relation between Anne's utterance of (1) and her nod that triggers the search for an appropriate entity from the nonlinguistic context. In answer to (ii) (*What is the semantic contribution of e and e′ to their respective discourses?*), we propose that the semantic contributions of e and e' are roughly like those of the italicized clauses in (5) and (6):

(5) Our daughter was sent to her room.$_{\pi_1}$ *She scratched up the wall.*$_{\pi_2}$ I was making dinner.$_{\pi_3}$

(6) I moved the table into the living room this morning and *I scratched up the wall*. I had to buy some new paint.

Accordingly, these contributions interact with the content of the surrounding discourse units in roughly the same way as the italicized sentences in (5) and (6) do. In (5), for instance, the content of the italicized sentence, labelled π_2 above, serves to explain that of sentence π_1, while π_3 provides background information regarding what was going on when the event described by π_2 took place. In other words, the content of π_2 affects the logical form of (5) by entering into two semantic relations: $\text{Explanation}(\pi_1, \pi_2)$ and $\text{Background}(\pi_2, \pi_3)$. In answer to question (iii) (*Where should the contributions of e and e′ be reflected in the logical forms for* Scratches *and* Table*?*), we propose that e affects the logical form of *Scratches* in a similar way, although π_2 should be replaced with a label ε that stands for the content assigned to e. The contribution of e' to *Table* can be analyzed along similar lines.

Our claim, then, is that e not only affects the interpretation of *Scratches*, but its very structure or logical form. Regardless of how one chooses to represent discourse structure — with trees, graphs, or stacks of discourse moves — it is impossible to build a representation of *Scratches* without countenancing a discourse unit that describes e. Given these observations, we can formulate a general hypothesis (NDU) about the way that e affects the interpretation of *Scratches* (and likewise for e' and *Table*):

> **The Nonlinguistic Discourse Unit Hypothesis (NDU)**: A nonlinguistic event e can affect the interpretation of a discourse by contributing the content of an entire discourse unit (i.e., an instance of a proposition), which enters into rhetorical relations with other,

> linguistically-specified discourse units. In so doing, e changes the very structure or logical form of the discourse, and it can do this without any explicit expression signalling its relevance. Rather, its relevance is inferred through the kind of reasoning used to infer rhetorical connections between linguistically expressed contents.

The main goal of this paper is to extend extant work on rhetorical structure to develop a model of situated discourse that takes (NDU) as a starting point. While (NDU) might seem very intuitive at first glance, it has nontrivial consequences for discourse structure and content that have not been explored in extant work. The sections that follow are dedicated to elucidating and modelling those consequences.

In pursuing (NDU), our work complements that of Lascarides & Stone (2009) and Stojnic et al. (2013), who have claimed that rhetorical relations play a key role in the semantic representation of multimodal interaction. Lascarides & Stone (2009) posit that coverbal hand gestures contribute discourse units to the logical form of a discourse, and they use rhetorical relations to model their contribution. Stojnic et al. (2013) allow nonlinguistic situations to contribute discourse units and posit a discourse relation *Summary* that connects linguistically-specified discourse units to situations (cf. topic situations in the work of Kamp (1981), Stanley & Szabó (2000), Elbourne (2005), among others). They use these connections to model how utterances can affect a salience ranking of nonlinguistic entities. Little is understood, however, about the range of coherence relations to which nonlinguistic events can contribute, or the nature and interpretation of the resulting structures. Our study expands upon this work by examining nonlinguistic events that are neither gestural nor coverbal and that are not necessarily salient in the way that topic situations are. We also consider a much wider range of interactions between nonlinguistic events and linguistic discourse. Finally, we examine the effects of multimodal interactions on global discourse interpretation and develop a perspective on discourse interpretation that has not, to our knowledge, been made explicit before.

2.2 From discourse to the interpretation of nonlinguistic events

A nonlinguistic eventuality that is observed or inferred from the nonlinguistic context cannot serve directly as an argument to a rhetorical relation; it must first be brought under a description or *conceptualization*. That is, it must be assigned a semantic content if it is to support inference. In general, however, there are many alternative conceptualizations of a nonlinguistic entity, be it an object or an event, and the linguistic context affects inferences about which of the alternatives is the most appropriate conceptualization on a given occasion. In *Scratches*, the scratches on the wall are understood as the outcome of an event of Peter and Anne's daughter

scratching up the wall, but in *Table*, we infer an event for which Anne is the agent. The nonlinguistic context remains constant, but the discourses as a whole support different conceptualizations of that context. Interpreters need a means for selecting a conceptualization that endows a nonlinguistic event with an appropriate, recursively structured content in a given context. Question (iv) (*How do e and e' come to be associated with the semantic content that they convey?*) inquires about the factors that determine this selection.

A consequence of the account that we will develop is that the content ultimately assigned to e and e' is inferred via rhetorical reasoning. That is, while nonlinguistic events can influence discourse content and interpretation by contributing content to discourse relations, information flows in the opposite direction as well: discourse structure influences the way that nonlinguistic events are interpreted or conceptualized. It is in trying to understand the coherent connections between Anne's utterances, her nod, and the visual environment that Peter will come to conceptualize e as an event of his daughter scratching up the wall. In fact, because e is not actually taking place at the time of utterance, coherence-based reasoning about the connection between Anne's signals and the wall must account for the inference that there is even a salient event that caused the (visible) scratch on the wall. Likewise for e'.

The co-dependence between the task of building discourse structure and the task of specifying the content of discourse units is well-known from work on purely linguistic discourse and rhetorical structure (Hobbs 1979). Moving to nonlinguistic events greatly complicates the latter task, however, because the typing information supplied by the nonlinguistic context is far more impoverished than that supplied by linguistic specification, and so the range of different possible interpretations (or conceptualizations) is greater.

Although we will make some remarks on the relation between discourse interpretation and conceptualization in this paper, a full answer to question (iv), as well as a full derivation of the content of either *Scratches* or *Table* is beyond its scope (though see the Appendix for a partial derivation). Providing such an answer would involve developing a model of discourse parsing that ties (probabilistic) models of perception of visual data with discourse information. Our main focus in this paper is to model the information flow from the nonlinguistic context to discourse structure and interpretation so that we have a solid foothold for studying conceptualization in the future. In other words, our goal is to develop the linguistic theory that could (and we think should) be used to inform a more sophisticated model of visual processing; one that draws on the coherence of embodied conversation as well as purely visual features. Our work thus complements that of Larsson (2013), Dobnik et al. (2013), and Zarrieß & Schlangen (2017), who investigate the conceptualization problem of nonlinguistic objects when a single clause or even a single word is uttered in a fixed

nonlinguistic context, in which the goals and interests of the agents are also clear and fixed.

If we are right that a successful classifier for visual information must rely on a theory of situated discourse of the sort that we develop in this paper, then our results will ultimately bear on the study of referential semantics as well.[1] A full story of reference to nonlinguistic entities will need to confront question (iv) — the conceptualization problem is by no means unique to our examples. Yet the fact that this problem rears its head even in the absence of any lexical triggers underscores our call for a more general model of conceptualization that looks beyond lexical semantics and composition. In our view, discourse structure is always a major factor in the conceptualization of the nonlinguistic context, whether or not this conceptualization serves as the interpretation of a lexical item.

3 A corpus for situated discourse

The semantic interactions between linguistic moves and nonlinguistic events that we aim to model take different forms and exhibit complexities beyond those we have discussed informally in connection with *Scratches* and *Table*. To facilitate the discussion of these complexities, we will exploit a corpus of chats taken from an online version of the game *The Settlers of Catan*, which we have annotated for rhetorical structure in the style of *Segmented Discourse Representation Theory* or SDRT (Asher & Lascarides 2003).[2] Our corpus has numerous advantages, which we will highlight shortly. First, however, we begin with some background on both *The Settlers of Catan* and the particular version of it that was used to build our corpus (for more details, see Asher et al. 2016).

The Settlers of Catan is a multi-party, win-lose game in which players use resources such as wood and sheep to build roads, settlements and cities on a game board. Players acquire resources in various ways, including trading with other players and rolling the dice. As shown in Figure 1, the game board is divided into hexes, each associated with a certain type of resource and a number between 2-6 or 8-12. A dice roll of, for example, a 4 and a 2 gives any player with a building on a hex marked "6" one or more resources associated with that hex. Rolling a 7 triggers a series of moves: the current player must move a game piece known as "the robber" to a hex of her choice and then steal a resource from a player with a building on that hex. The robber will then stay on that hex until moved in another turn, and its

1 While we do not have space to elaborate on this point here, an account of how discourse relations influence the interpretation of referentially used expressions would complement work by Andy Kehler and colleagues that considers how discourse relations influence the interpretation of anaphoric expressions (Kehler 2002).

2 To view the annotations for the corpus, go to https://www.irit.fr/STAC/corpus.html.

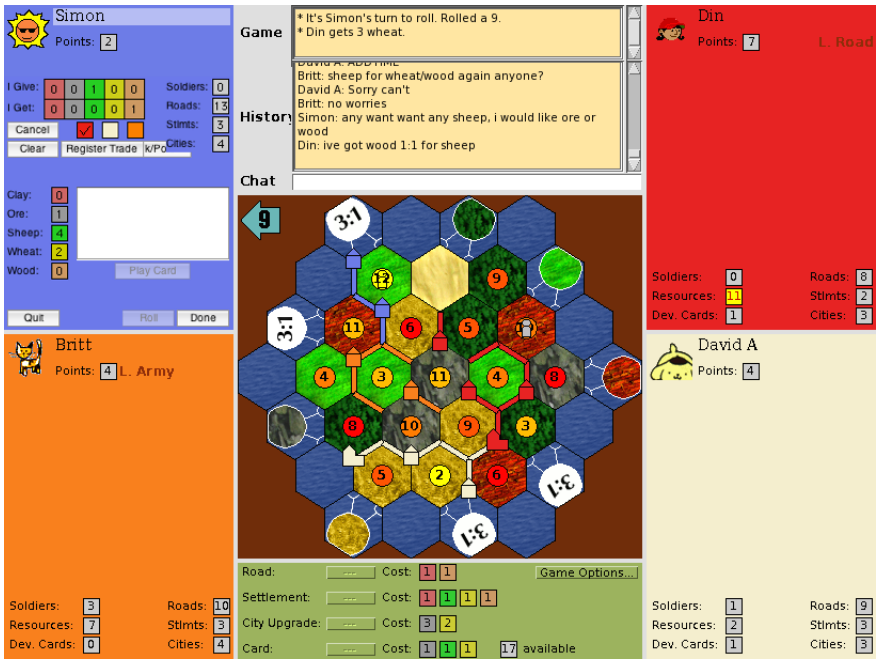

Figure 1: A snapshot of the game board for *The Settlers of Catan*

presence will continue to impact the game by blocking resource distributions for the occupied hex, regardless of the dice rolls.

Figure 1 illustrates the game interface that was used to collect our corpus. This particular snapshot is the way the game board looks to the player Simon. We can see Simon's resources in the upper left-hand rectangle of the game interface, but as in the physical version of the game, Simon cannot see the resources of his opponents. Figure 1 also shows that Simon is preparing to make a trade via a *Trade Panel*: he has prepared an offer to the (red) player Din, but has not yet clicked "Register Trade". The window labelled *Game* at the top of Figure 1 reflects game events that are public to all players. Once Simon registers his trade, for instance, then Din's response, whether he accepts or rejects the trade, will be described in the Game window.

To construct our *Settlers* corpus, a chat option was added to the online game interface, as shown in Figure 1 (Afantenos et al. 2012). A player types in the *Chat* window, and prior chat is recorded in the *History* window. To encourage discussion, players were instructed to negotiate trades in the chat interface before executing an agreed trade through the Trade Panel. Interestingly, for the purposes of this paper, players took advantage of the chat window to chat not only about trades, but about many aspects of the game state. (8) is one example:

(8)

433.0.3	Server	william played a Monopoly card.
433.0.4	Server	william monopolized wheat.
433.0.5	Server	It's william's turn to roll the dice.
434	GWFS	noooo!
435	Server	william rolled a 2 and a 1.
436	Server	GWFS gets 1 sheep. LJAY gets 2 wood. T.K. gets 2 wood.
436.0.0.1	UI	GWFS has 4 resources. LJAY has 3 resources. william has 13 resources.
437	GWFS	greedy :D
438	william	:D
439	GWFS	spend it wisely then
440	LJAY	:)
441	LJAY	13! :o

Every turn in our corpus, whether it is a chat move or a game event, is assigned a turn number; the turn numbers for (8) are indicated in the left column.[3] Each turn is also identified with an agent, as shown in the middle column of (8). For chat moves, the agent is the player who typed the chat message (e.g., GWFS for turn 434);[4] game events and states are either described in Server messages, many of which were visible to all players in the Game window, or reconstructed (by our team) using information from the User Interface (UI). In (8), William plays a Monopoly card, which allows him to steal all instances of a particular resource of his choice that are possessed by the other players. In turn 433.0.4, he steals all of the wheat. Both GWFS and LJAY comment on William's move. There is some ambiguity as to whether LJAY in 440 comments on the theft itself or on GWFS' comments in 437. Immediately after, in 441, LJAY comments on the result of the theft: that William has 13 resources.

(8) illustrates clearly the point that we made in Section 2.1 that building a complete and correct representation of a situated discourse can force us to countenance discourse units contributed by nonlinguistic events. We cannot build a discourse graph that accurately reflects the connections accounting for the coherence of (8) using only the moves 434, 437, 438, 439, 440, and 441. In fact, because the *Settlers* corpus was not originally created to study situated discourse, annotators were initially given only the chat moves to annotate for discourse structure. The resulting annotations were often incomplete — annotators could not find a reasonable point of attachment for many chat moves, and so left them as 'orphans' (i.e., disconnected) in the discourse structure. Turn 434 in (8) was such an orphan.

3 Annotations of the *Settlers* corpus have taken place over multiple stages. Game messages that were added in a later stage are given decimal numbers in order to preserve the original numbering of the chat and game events that were present in the first stage.

4 Small capitals indicate user names that have been abbreviated to save space or preserve anonymity.

These observations triggered a second round of annotations in which annotators had access to both chat moves and game events, and a comparison of the two rounds of annotation has allowed us to quantify the incompleteness of the original discourse structures. The results show, for example, a complete reduction in orphaned discourse units in the second round of annotations. (The first set of annotations contained 364 dialogue internal orphans, where dialogue boundaries were principally determined by dice rolls,[5] and 1501 orphans total out of a total of 12588 linguistic moves in the 45 games of the corpus.) While this comparison reflects features specific to the *Settlers* corpus, it also has the general effect of highlighting the extent to which nonlinguistic events can contribute directly to discourse structure.

The inability to build complete discourse representations without the game moves is not the only issue. The comparison between the two sets of annotations also shows that we cannot even reliably reconstruct the partial structures to which the linguistic moves contribute. This is because the game events affect inferences about which relations hold between even the linguistic-only moves. In more detail, 2006 of the 9879 discourse connections in the first round of annotations, which were posited based on linguistic-only information, were judged incorrect once annotators gained access to game moves (for more details, see Hunter et al. 2015). In addition, we found that adding the game events affected judgments about the way that chat-only moves are grouped together. Sometimes, a group of discourse units work together to provide a single, collective argument to a discourse relation. We found that about 10% of the 1450 groupings of chat moves in the chat-only annotations either changed or disappeared, and about 34% new groupings of chat moves (only) were added as a result of taking the nonlinguistic context into account. (9) illustrates a common pattern in which linguistic moves are grouped together because they jointly cause a nonlinguistic event.

(9)

71	T.K.	anyone can offer any wood?
72	william	sry no
73	GWFS	sorry - more 6s and I can oblige then :)
74	LJAY	move the robber and sure :p
75	Server	T.K. traded 3 sheep for 1 wood from a port.
76	Server	T.K. built a road.

Intuitively, neither 72, 73, nor 74 is *independently* responsible for T.K.'s decision to trade from a port.[6] 75 is rather the result of his entire failed attempt to trade with the other players; that is, the first argument of the inferred Result to 75 is a *complex*

5 When there were discourse relations that linked moves across dice rolls, we opted to treat these contributions as one dialogue.

6 A port is a particular location on the Catan board where players can get preferential trades under certain circumstances.

discourse unit composed of 71-74.[7] The first graph below illustrates the structure inferred from the first round of annotations; the second graph illustrates the new structure that groups moves 71-74 together into a complex discourse unit, indicated by the box, so that it serves as the first argument to the Result relation. QAP stands for the relation Question-Answer-Pair.

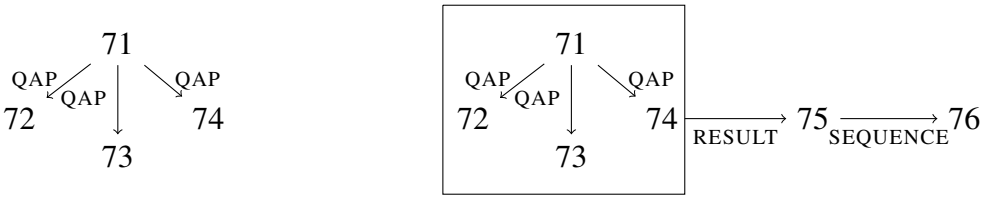

Figure 2: The original graph for chat moves 71-74 (left) and the updated graph, in which 71-74 form a complex discourse unit (right).

(8) also highlights a great advantage of our corpus: the Server and UI messages allow us to disentangle the co-dependent tasks involved in more natural examples like *Scratches* and *Table* because they almost always settle the conceptualization of the basic game events (see Section 2.2).[8] This makes our corpus particularly interesting for researchers in computational linguistics and human-robot interaction, who wish to empirically study semantic dependence on nonlinguistic events. One of the biggest hurdles of such a study comes from the fact that the nonlinguistic context often involves a seamless evolution of perceptual input, yet to study the semantic effects of the nonlinguistic context, we must be able to individuate semantically relevant events and objects, and we must be able to assign them an appropriate typing or interpretation.

The fact that our *Settlers* corpus largely circumvents the conceptualization problem is partly due to its game-centered environment: even in a physically embodied version of the *Settlers of Catan* board game, the mutually known rules of the game enable observers to individuate events like dice rolls and card plays, and to describe them in terms of their role in game play. But in a physically embodied environment, there would also likely be other linguistic/nonlinguistic interactions that would be

7 Asher et al. 2016 describes the SDRT-inspired annotation of the corpus. For more on SDRT representations see Section 5.1 and the Appendix.

8 There are some limited exceptions. For example, players can misconceive to whom they are directing a trade offer. When setting up trades, players are identified by colors, not their names, so such mistakes in conceptualization occasionally happen. There are also cases in which a sequence of chat moves can affect the interpretation of game events, as in example (12) in Section 4.2.

harder to control. The virtual environment of our *Settlers* game avoids these complications. In addition, for every game in our corpus, game events and states are either already described by Server messages, or easily recovered from UI information, meaning that individuation and typing of these events is almost entirely straightforward.

Of course, one might worry that because the Server and UI yield descriptions, the game events in our corpus should not be counted as nonlinguistic events, meaning that our corpus does not, in fact, shed light on information flow from the nonlinguistic context to discourse interpretation. We do not think this is a concern. First of all, the Server messages that a player sees often fail to fully specify a game event. The location of the robber, for instance, is never verbally presented. When a player moves the robber, the Server message broadcast to the players in the Game window is *[player i] has moved the robber*, but this does not specify *where* it has moved: the players must perceive that on the game board. In fact, UI information, such as who is sitting where, when a turn is ended, and to whom a Trade Panel offer is made, is given only visually to the players, rather than encoded in Server messages in their Game windows; and players regularly engage with this type of information. The UI descriptions found in our examples were reconstructed by our team after the corpus was collected, using information in the game logs.

Moreover, the players do not need to rely on the Server messages to know what is going on in the game any more than a player needs to rely on a sportscaster to know what is going on in a football game. The messages are helpful as a record for annotators and for players who might have a lapse in attention, but all of the information conveyed by the Server is represented visually to the players. Players can tell when the dice have been passed to a new player because a pointer on the screen moves to the part of the screen dedicated to that player; other players can tell when the Red player gets a resource via the visualised dice rolls and where the Red player has built settlements; and so on. The fact that the Server linguistically describes the game events might pose a problem if we were studying conceptualization, but as we stated earlier, we are concerned in this paper with information flow in the opposite direction. We want to know how allowing events such as dice rolls and card plays to contribute discourse units in a semantic representation of an embodied conversation affects the development and interpretation of a discourse structure, and our corpus facilitates this study even if on occasion the players learn about the events via the Server messages.

This brings out a deeper point about the nature of the game events: players can interact with them in the same way regardless of whether or not they are associated with speech acts. This suggests that the distinction between nonlinguistic and linguistic events is not the only one relevant to the interpretation of situated, multimodal discourse. What is interesting about the game events in (8) is that while

we understand (8) as an integrated, coherent discourse, the game events actually form a substructure whose nature is largely independent of the content of the chat moves. Thus while these events happen to contribute to the content of (8) via their coherence relations to linguistic discourse units, this is not their *raison d'être*. The agents clearly do not make their game moves with the intention that they contribute to the content of a discourse, but rather with the intention of winning the game. Nevertheless, the speakers exploit the game events and *appropriate* them for their discourse purposes, making the chat moves parasitic on the game structure (regardless of whether we classify the game events as linguistic or nonlinguistic).

4 Conceptual shifts

Developing an adequate model of examples like *Scratches*, *Table*, and similar examples from our corpus requires rethinking the status of nonlinguistic events and reexamining standard assumptions about coherence relations. In Section 4.1 we introduce a distinction between discourse dependent and discourse independent events. This is relevant for studying situated discourse, but we also explain that it cuts across the linguistic/nonlinguistic distinction. In Section 4.2, we apply the discussion of discourse independent events and multimodal interactions to an old discussion about the role of intentions in interpretation.

4.1 Discourse independent and discourse dependent events

Like most activities, conversations are typically goal directed. Though such goals might be hard to articulate and we will leave this point implicit here, we take the intuition to be clear.[9] As with most activities, some events that occur during the course of a conversation will occur in order to further the conversation's goal; other events might simply happen to occur at the same time, possibly contributing to some other goal, independent of that of the conversation. In a situated discourse d, for example, cars whose drivers are unaware of d might pass on the streets; a busker who hasn't noticed d's participants might start playing music in the background; a group of people walking by the participants of d might be carrying on their own conversation, oblivious to the conversations of others in the situation, and so on. Such events are clearly *discourse independent* events; they take place in the context of d, or may have effects that are apparent in that context, but they are not constrained to be relevant to d. A *discourse dependent* event e, on the other hand, takes place in order to further the main goals of d; had the goals of d been different, e would not have occurred, or at least would not have entered into the same discourse connections.

9 For our theory of goals, see Asher et al. (2017).

Sometimes, conversational participants will exploit discourse independent events and incorporate them into their discourse, effectively turning them into discourse moves. In this case, the content, structure, and goals of the conversation will depend on the discourse independent events in the same way that they depend on discourse dependent moves; at the same time, the discourse independent events will still be intuitively understood as occurring for reasons independent of the conversation. This kind of asymmetric dependency — in which a discourse depends on discourse moves whose existence and interpretation do not themselves depend on the discourse — is common in the *Settlers* corpus. Let's return to (8):

(8)

433.0.3	Server	william played a Monopoly card.
433.0.4	Server	william monopolized wheat.
433.0.5	Server	It's william's turn to roll the dice.
434	GWFS	noooo!
435	Server	william rolled a 2 and a 1.
436	Server	GWFS gets 1 sheep. LJAY gets 2 wood. T.K. gets 2 wood.
436.0.0.1	UI	GWFS has 4 resources. LJAY has 3 resources. william has 13 resources.
437	GWFS	greedy :D
438	william	:D
439	GWFS	spend it wisely then
440	LJAY	:)
441	LJAY	13! :o

In (8), the game moves form a kind of independent structure of their own in the sense that were we to ignore the chat moves, the structural connections between the game moves would remain intact, and their interpretation, unchanged — the *raison d'être* of the game moves is independent of the accompanying discourse. By contrast, the chat moves depend on the game moves in the sense that were we to ignore the game events (as we in fact did in the first round of annotations on the *Settlers* corpus), then we could infer different structural connections between certain chat moves while other chat moves would be left "orphaned" as anaphoric fragments without antecedents. The *raison d'être* of the chat moves is ultimately to interact in a coherent fashion with certain game moves.

The structure S determined by the whole of the interaction in (8) exemplifies what we will call an *asymmetric dependency*. S contains a *core* C of connected, independently interpreted moves that can be distinguished from moves in what we will call the *periphery* P of S. The asymmetry of S derives from the fact that the moves in p depend on the connected elements of C in the sense that they presuppose them for their coherence while C does not presuppose the existence of structures

in *P* to form a coherent structure. A result of this asymmetric dependence is that while *C* will be a connected substructure of *S*, the moves in *P* need not form a single connected substructure of *S* but rather a set of connected substructures, each one connected to some element of the core *C*. Accordingly, we can characterize *P* moves as "appropriating" elements in *C*, while elements of *C* do not appropriate any part of *P*.

While the core of (8) contains all and only the game moves in (8) and its periphery contains all and only the chat moves, the distinction between asymmetric structures and other structures actually cuts across the linguistic/nonlinguistic distinction. For one thing, the core of an asymmetric structure could contain both linguistic and nonlinguistic moves, thereby forming an *interleaved structure* — a notion that we introduce below. Asymmetric dependencies are moreover not restricted to multimodal interactions — conversational participants who overhear another conversation or listen to a public speech, for example, might appropriate linguistic moves from the external discourse and incorporate them into their own conversation leading to a fully linguistic asymmetric structure. An asymmetric dependency is not characterized by the types of moves that figure in its core and periphery, but by the ways in which the moves affect the development of discourse structure.

Another type of structure manifested by multimodal discourse in which a nonlinguistic event can contribute a whole discourse unit is what we will call an *interleaved structure*. Interleaved structures feature units derived from linguistic and nonlinguistic material that function together on an "equal" basis, contributing together to causal, narrative, and dialogical (e.g., question answer pair) substructures to the discourse structure. One frequent case in our corpus is when a trade negotiation leads to a nonlinguistic trade, as in (9), repeated below.

(9)

71	T.K.	anyone can offer any wood?
72	william	sry no
73	GWFS	sorry - more 6s and I can oblige then :)
74	LJAY	move the robber and sure :p
75	Server	T.K. traded 3 sheep for 1 wood from a port.
76	Server	T.K. built a road.

In this structure, as opposed to the asymmetrically dependent structure in (8), the chat moves bring about a change in the game state.

The connections between nonlinguistic events and linguistic moves in an interleaved structure are in many ways like those familiar from purely linguistic discourse. (10), however, provides another, more subtle example that shows that interleaved structures are nevertheless different from structures for purely linguistic discourse, at least as the latter are normally conceived.

(10)

534	GWFS	anyone want to trade their ore for my wood?
535	LJAY	nope
538	GWFS	it may prove a prudent trade, lj…
539	LJAY	nope
539.1	Server	GWFS played a Soldier card.
539.4	Server	GWFS stole a resource from LJAY
540	GWFS	apologies…
541	LJAY	:(

In this example, GWFS has tried, but failed, to trade with LJAY. Despite the warning in 538 that it might be a good idea for LJAY to trade, she rejects his trade offer (again) in 539. GWFS reacts by playing a Soldier card, which allows him to steal from her. LJAY then expresses her disappointment in 541.

A reasonable follow up question to GWFS's warning in 538 would be *Why?* or *Why would it be prudent?*. The moves 539.1 and 539.4 answer this question; they *explain* why he said 538. The fact that GWFS plays a Soldier card shows that his stealing from LJAY was a planned attack, and after he carries out the robbery, we come to understand that he was not merely telling LJAY that the trade might pay off for her in the long run, but giving her a specific warning in light of his chosen backup plan. Reasoning about the connection between 538 and 539.4 helps us to interpret 538.

At the same time, GWFS did not play the Soldier card and steal a resource from LJAY *in order to explain* his warning any more than Peter and Anne's daughter scratched up the wall in order to explain why she was sent to her room. GWFS made his moves in order to get a resource from LJAY and to place himself in a better position to score a victory point in the game. The *raison d'être* of the warning was to encourage the development of the game in one direction rather than another; as such, the warning is internal to the game structure, which yields an interleaved structure. At the same time, the explanation that we infer, which is a meta-level relation concerning GWFS's strategy, is external to the game structure. Thus even internal moves in an interleaved structure can give rise to asymmetric dependencies, when these moves play a kind of double role. So to be precise, when we speak of a discourse unit as being part of the core or the periphery, we mean the element *as it plays a particular role*.

4.2 The hermeneutical stance

In the previous section, we saw how Hypothesis (NDU) and data like (8) and (10) lead to the existence of asymmetric dependencies and interleaved structures. In this section, we consider how these structures impact basic assumptions about the nature

of discourse relations and structure. We take it to be a standard assumption that if the content p of a discourse move m stands in an Explanation relation to the content q of a discourse move n such that p provides the explanans, then the *raison d'être* of m is to provide an explanation of q. That is the function of m, and that is why p was added to the content of the discourse. Discourse independent events that are appropriated in discourse as well as the nonlinguistic events in interleaved structures undermine this assumption by contributing arguments to rhetorical relations to which it is not their *raison d'être* to contribute.

This has consequences for an old debate about the role of communicative intentions in discourse interpretation. On one side of this debate are Griceans who hold that communicative intentions are constitutive of interpretation: for an interpreter to infer an Explanation between p and q, she must recognize that the speaker expressed p with the intention of using p to explain q and she intended for this intention to be recognized. Thus the reasoning in interpretation flows from inferences about intentions to inferences about content. On the other side of the debate are those who ascribe a less central role to communicative intentions: an Explanation can be inferred on the basis of features of p and q and from there, an interpreter can defeasibly infer that the speaker had the intention of using p to explain q (Lepore & Stone 2015, Asher et al. 2017). In other words, the reasoning flows the other way, from a preferred pragmatic interpretation to intentions.

When it comes to nonlinguistic, discourse independent events and nonlinguistic events in interleaved structures, the events do not arise from an intention to communicate. Their *raison d'être* is to make things happen in the world. At the same time, we have argued that these events contribute to a discourse structure (or game structure, etc.) in *the very same way* as discourse moves do. We need this claim in order to explain the coherence of an example such as (10): the chat moves in 538, 539, 540, and 541 are intuitively a part of a connected and coherent interaction, but representing the coherent connections requires the game moves to contribute to the representation of (10)'s structure. These claims together render moot the question of whether communicative intentions enter the picture before or after the inference to a coherence relation, because the requisite events, and in many cases the inferred relations to which they contribute, are simply not produced from an intention to communicate.

In our view, semantic structures composed entirely of what are traditionally classified as discourse moves (including, perhaps, discourse dependent nonlinguistic moves) are just a subclass of the kinds of structures that we can use such moves to build. In fact, we think that the kinds of semantic structures built up from coherence relations need not involve any discourse moves at all. Suppose Peter looks out into the garden and sees his cat, Lupin, staring at a pile of leaves. The leaves suddenly move, and Lupin pounces. Peter goes to investigate and finds a baby whipsnake. He

now understands why Lupin was staring at the leaves and why the leaves rustled; he also understands that Lupin's pounce was a result of the leaf movement. Yet, neither the snake nor the cat intended to communicate anything, and certainly the snake didn't intend its presence to explain Lupin's behaviour and Lupin didn't intend his pounce to be a result of the leaf movement. Nevertheless, both the result and meta-level explanation are inferred.

Interpreting the asymmetric and interleaved structures in our *Settlers* corpus often requires inferring coherence relations between game events. In (11), the distribution of resources in 205 is a result of the dice roll in 204.

(11)

204	Server	J rolled a 2 and a 3.
205	Server	mmatrtajova gets 1 sheep. Ash gets 1 sheep.
206	mmatrtajova	nicee
207	J	my dice rolls SUCK

J's comment in 207 brings not only the moves 204 and 205 into a larger semantic structure, but the relation between them as well: J's dice roll sucked precisely because it *resulted* in a resource distribution for her opponents while yielding nothing for her. In other words, by Hypothesis (NDU), 207 is coherently related to a complex unit ε, whose content is Result(204,205). But communicative intentions are not responsible for ε's content; indeed, *no* intentions are, at least not directly. 204, 205, and their causal relationship stem from the rules of the *Settlers* game, but 207 makes this Result relation semantically relevant in a way that requires us to interpret it as contributing to a larger, asymmetric structure.

When discourse independent events get appropriated and incorporated into a larger discourse structure, this process can lead to a re-conceptualization of game events analogous to the regrouping of chat events that we observed in (9). Nevertheless, the substructure retains its independence in the sense that peripheral moves will not lead us to correct connections inferred on the basis of core moves alone; rather, one only groups core moves together in new ways. Consider (12):

(12)

154.1	Server	GWFS played a Soldier card.
154.3	Server	GWFS stole a resource from LJAY
155	Server	GWFS rolled a 5 and a 1.
157	Server	GWFS built a settlement.
158	GWFS	sorry laura
159	GWFS	needed clay the mean way :D
159.1	Server	LJAY played a Soldier card.
159.4	Server	LJAY stole a resource from GWFS
160	Server	LJAY rolled a 4 and a 4.
161	Server	GWFS gets 2 wheat.
163	GWFS	touché

When GWFS types *touché*, he comments on the fact that LJAY made a successful counter move — rather than accepting his apology she "retorted" by attacking him back. Interpreting GWFS's comment in this way therefore requires that we infer a parallel structure between LJAY's complex play in 159.1-159.4 and GWFS's prior complex move in 154.1-154.3. In inferring this connection between complex discourse units, we add information to the representation of (12) that goes beyond what is strictly required by the game structure alone. Still, the substructure of game events — what we have called the core — yields a complete discourse structure C that is consistent with this reconceptualization.

The need to infer semantic connections between real-world, noncommunicative events has general consequences for the way we think of discourse. Dynamic semantic theories of discourse interpretation are based on the assumption that discourse is fundamentally about information exchange, and so concentrate on how speakers convey information, looking at questions such as: How does the content of a given utterance affect the information state(s) of conversational participants? How does it restrict the set of possible subsequent discourse contributions (e.g., how does it affect salience)? And how does the packaging of information in a given utterance serve to further a larger, information-seeking discourse goal? Once we expand our view to the full set of interactions with real-world events, however, we see that the set of questions that we must pursue, and thus the models of discourse evolution and interpretation that we develop to answer them, must change significantly.

Discourse evolution, for example, is constrained by the salience of discourse moves — for a new discourse move to be coherent, it must typically relate to a salient prior discourse move (if it is not discourse initial). Once we countenance asymmetric structures, we must consider not only the salience of discourse moves, but the salience of real-world events, which may very well be unfolding according to discourse independent rules and goals. We therefore have to consider not only how a discourse move can interact with an isolated nonlinguistic event, but with an entire *structure* of nonlinguistic events (or multimodal interactions). Discourse salience cannot be understood only in terms of the production of discourse moves, but also in terms of how speakers can exploit discourse external events. We explore these issues in more detail in Section 5.1.

Discourse interpretation likewise needs to take real-world events into account. There is an illustrative analogy with deixis: the interpretation of a deictic expression is determined ultimately by the way the world is, not the way that a speaker describes it as being, and its interpretation does not vary. A speaker might be wrong about what she takes the referent of a deictic expression to be, but the real world has the final say as to what the referent is. Similarly, the introduction of a variable and associated content for a nonlinguistic event characterizes or describes something that actually happens in the world. Once that nonlinguistic event is introduced, it determines an

intensional content that does not vary with circumstances of evaluation and that escapes the scope of operators such as negation or modality. It is this intensional content that various conceptualizations may or may not correctly capture, that speakers may get wrong or right. But in an important sense, when speakers interact with a nonlinguistic event in an asymmetric or interleaved structure, they become committed to the event itself, not only to a particular conceptualization of it. And the event, regardless of whether conversational participants realize it or not, can limit possible discourse continuations and force a discourse to take a turn in a way that no amount of discussion can change. We develop this point in Section 5.2.

The interpretation of deictic expressions has traditionally been handled using mechanisms that are external to dynamic discourse interpretation: Discourse Representation Theory, for example, uses external anchors to fix a unique assignment function for deictic expressions (Kamp 1990). Kaplan's (1989) model of deixis is importantly similar: the referents of deictic expressions are fixed antecedently by a special *character* function, and then discourse interpretation can proceed as standardly construed. Hunter (2014) and Maier (2009) have already argued for a more integrated model of deixis, but the need to take real-world effects into account during dynamic discourse update itself is even clearer in the kinds of exchanges that we are considering. Simply adding anchoring or character functions is not an option in these cases, for there is no expression to anchor. The events are integral to the very construction of the discourse.

5 Technical changes to discourse structures

We have argued that building discourse structures for situated discourse is not as simple as adding nonlinguistic events to our model and allowing the interpretation of our discourse structures to be sensitive to them; we must allow these events to contribute contents directly to the representations. Extending our structures in this way leads us to countenance asymmetric and interleaved discourse structures, which, as we argue in this section, require us to revise standard models of discourse structure and interpretation. In Section 5.1, we examine the consequences of asymmetric structures for discourse structure and evolution. Section 5.2 explores the ways in which incorporating nonlinguistic events into discourse content, be it in interleaved or asymmetric structures, affects dynamic semantic models of discourse.

To make our discussion more concrete, we will adopt some of the language and formalisms of *Segmented Discourse Representation Theory* or SDRT (Asher & Lascarides 2003). Few of the general points that we make in this section depend on details specific to SDRT, however, though they do rest on the assumption that the content and structure of a coherent discourse can be represented as a weakly connected graph that permits long distance dependencies; the consequence of this

is that a discourse move is coherent in the context of a discourse only insofar as it is coherently related to some other (potentially nonadjacent) part of that discourse. Details of the formal implementation in SDRT can be found in the Appendix.

5.1 Structural constraints on situated discourse structures

Asymmetric structures are formed when discourse moves interact with an independently developing structure to form a coherent whole. This interaction often involves discourse units playing the double role of engaging in connections with other discourse moves while simultaneously engaging with elements of the independently developing structure, which we have called the *core*. This double role in turn requires that we rethink constraints on discourse salience and development: whereas these constraints are normally modelled in terms of how a speaker should formulate and present her discourse contributions, asymmetric structures lead us to ask how a speaker can exploit events in an independently evolving structure to achieve her discourse goals. To spell out these changes, we begin in Section 5.1.1 by defining asymmetric structures, including the notion of a core and a periphery, more precisely. Section 5.1.2 then describes two patterns of discourse attachment exhibited by asymmetric structures in the *Settlers* corpus that are disallowed by extant theories of discourse. Finally, in Section 5.1.3, we propose a new constraint on discourse salience and evolution that takes these new patterns of attachment into account.

5.1.1 Asymmetric structures

An asymmetric structure contains a substructure that has an autonomous existence, its core, and one or more substructures that are dependent on the core. This set of dependent structures makes up what we have called the *periphery*. To define asymmetric structures precisely, we therefore need to define the core of a structure and its periphery. We begin by defining a discourse graph in SDRT[10] — a weakly connected graph with directed edges or arrows that are labelled with names for discourse relations (e.g., Contrast, Elaboration, Result, Commentary, among others). The nodes in our graphs will be either: (i) *elementary discourse units* (EDUs), which we will take to be the contents of linguistically-specified clauses; or (ii) *elementary event units* (EEUs), which are contents assigned to nonlinguistic events in the context; or (iii) *complex discourse units* (CDUs). A CDU is a discourse unit whose content constitutes coherently connected discourse sub-units of types (i), (ii) and/or (iii).

10 The *Settlers* corpus satisfies the presumption in Definition 1 that the units are in a total linear order, but multimodal conversation is generally nonlinear. We eschew this complexity here.

Definition 1 *A* **discourse graph** G *is a tuple* $(\mathsf{V}, \mathsf{E}_1, \mathsf{E}_2, \ell, \mathsf{Last})$*, where* V *is a set of nodes* (EDUs, EEUs, *and* CDUs)*;* E_1*, a set of edges representing discourse relations;* E_2*, a set of edges relating each complex discourse unit* (CDU) *to its sub-units;* ℓ*, a labelling function from elements of* E_1 *to discourse relation types; and* Last*, a label for the last unit in* V *relative to textual order.*

The *core* of an asymmetric structure can now be defined as a particular kind of subgraph. For a graph G, a core C will be a subgraph G' of G that consists of a chain of edges that connect the first element of G to the last element of G.

Definition 2 *Let* $\mathsf{G} = (\mathsf{V}, \mathsf{E}_1, \mathsf{E}_2, \ell, \mathsf{Last})$ *be a discourse graph and let i be the initial* DU *in* V *with respect to the textual order. A subgraph* $\mathsf{G}' = (\mathsf{V}', \mathsf{E}'_1, \mathsf{E}'_2, \ell^{\mathsf{G}} \upharpoonright \mathsf{E}'_1, \mathsf{Last})$ *of* G *forms a* **core** C *just in case: (i)* $\{i, \mathsf{Last}\} \subseteq \mathsf{V}'$*; (ii) the transitive closure of* E'_1 *induces a transitive, asymmetric ordering R over* V' *in which for every element a, other than* Last *and i,* $R(i,a)$ *and* $R(a, \mathsf{Last})$*.*

Note that any maximal chain over V, defined in the standard way, is a core, and any set of chains over V all with the same endpoints forms a core as well.

When a graph G represents an asymmetric structure, the *periphery* of the structure is the set of edges that remain when we remove a core from the graph. To make this precise, Definition 3 first defines the set of edges that is left in a graph G when a subgraph G' is removed from G. Definition 4 makes such a remainder a *periphery* when the subgraph that is removed is a core C. Definition 5 entails that an asymmetric structure is a graph with a core and a nonempty periphery.

Definition 3 *Let* G *be a discourse graph, and* G' *a subgraph of* G*. Let* $End(e)$ *be the endpoints of an edge e and* $End(\mathsf{E}) = \{x\colon \exists e \in \mathsf{E}. x \in End(e)\}$*. Let* $\setminus$ *stand for set-theoretic difference, and let* $End(\mathsf{E}_1 \setminus \mathsf{E}'_1) = \mathsf{V}^{\mathsf{P}}$ *(for* periphery*). Then* $\mathsf{G} - \mathsf{G}' =_{defn}$ $(\mathsf{V}^{\mathsf{P}},\ \mathsf{E}_1 \setminus \mathsf{E}'_1,\ \mathsf{E}_2 \upharpoonright \mathsf{V}^{\mathsf{P}},\ \ell \upharpoonright (\mathsf{E}_1 \setminus \mathsf{E}'_1),\ x)$*, with x the last element in* $\mathsf{V} \setminus \mathsf{V}'$ *ordered by a linear ordering* $\prec$ *over* V*.*

Note that $\mathsf{G} - \mathsf{G}'$ may not be a discourse graph in our sense in that it is no longer weakly connected; nevertheless, it corresponds to a set of discourse graphs. We will assume this correspondence below. Note also that for a given discourse graph G and substructure G', G' and $\mathsf{G} - \mathsf{G}'$ may share nodes but form a partition over the set of relation instances or arcs in E_1.

Definition 4 $\mathsf{P}(\mathsf{G}, \mathsf{C})$*, the* **periphery** *of a structure* G *with respect to a core* C*, is such that* $\mathsf{P}(\mathsf{G}, \mathsf{C}) = \mathsf{G} - \mathsf{C}$*.*

Definition 5 *An* **asymmetric structure** G *is a graph with a core* C *such that* $\mathsf{C} \neq \mathsf{G}$*.*

93.0.2	Server	It's Dave's turn to roll the dice.
94	Server	Dave rolled a 3 and a 4.
94.0.1	Server	rennoc1 needs to discard.
95	Dave	:D
95.0.1	Server	rennoc1 discarded 4 resources.
95.0.2	Server	Dave will move the robber.
95.0.3	Server	Dave stole a resource from Tomm
96	Tomm	[Oh. . .] [well that's unkind]
97	Server	Dave built a road.
98	Dave	sorry
99	Tomm	[It's okay,] [name of the game and all that.]
100	Tomm	Do you get a random one of my resources then?
101	Dave	anyone need wheat?
102	Dave	I did, yeah
103	Tomm	no thanks
104	Dave	rennoc?
105	rennoc1	i'm ok for now
105.0.0.1	Server	Dave bought a development card.
105.0.1	UI	Dave ended their turn.

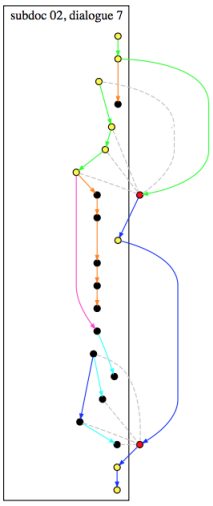

Figure 3: A dialogue and its asymmetric structure. The core consists of all the yellow EEU nodes plus the four black EDU nodes contained in the bottom red CDU node.

A core in an asymmetric structure G is like the backbone of G; the periphery on its own typically does not form a connected graph. Figure 3 illustrates an asymmetric discourse structure, representing the content of an entire situated dialogue. The black nodes represent EDUs; the yellow nodes, EEUs. The two red nodes represent CDUs and the dashed edges extending from them indicate their members. The top CDU (red node) contains the four EEUs (yellow nodes) that result from Dave's rolling a 7 (so 94.0.1 and 95.0.1-95.0.3). The bottom CDU contains the EDUs (black nodes) 101, 103, 104 and 105 that constitute a failed bargaining attempt. The core consists of all the yellow EEU nodes plus the four black nodes contained in the bottom CDU.

The periphery in Figure 3 contains two structures. The first is a tree of depth one that links the EDU 95 to the EEU 94 in the core; the orange arc indicates that 95 is a commentary on 94. The second structure in the periphery is a more complex tree hanging off EEU 95.0.3, the last EEU figuring in the top CDU. A pink arc, which signifies a clarification question, links 95.0.3 with 100; the blue arc following the pink arc indicates that the question was answered by EDU 102. The line of orange arcs extending from 95.0.3 signifies a series of commentaries made in the two EDUs in turn 96 (indicated by brackets), 98, and the two EDUs in 99. Taking the periphery away from the structure leaves an intact connected structure, while taking away the core from the whole structure would lose connectedness.

Figure 3 illustrates an important structural characteristic of asymmetric structures that helps us relate them to other, known kinds of discourse structure in SDRT. The periphery of an asymmetric structure with a maximal core is distinguished by the

fact that there are no outgoing arrows from elements of the periphery to elements of the core. Theories like SDRT already countenance the possibility of outgoing arrows that extend from an element of a CDU but which do not play a central role in the progression of a discourse; SDRT calls such outgoing edges, or subgraphs built from them, *danglers* (Venant et al. 2013). Units contributed by appositive relative clauses, for example, generally function as danglers even in discourse representations of single authored text. The periphery of an asymmetric structure thus resembles a collection of one or more danglers. However, while we do not have the space to develop this point here, danglers provided by appositive relative clauses or presuppositional constructions typically interact differently with content in the core. Speakers often use them to provide more information about an entity under discussion, and so the appositive relative clause helps an interpreter to situate discourse content in a way that the commentary in our corpus generally does not.

To make the notion of a dangler more precise, we first define a maximal core. As the edges in an SDRT graph are directed, the transitive closure of the set E_1 will yield an asymmetric ordering with a maximal element, namely the first element in the ordering. Let us call a core C of a graph G *maximal* just in case there is no substructure A of G such that $\mathsf{P}(\mathsf{A},\mathsf{C}) \neq \emptyset$ and A is also a core of G.

Fact 1 *Every weakly connected directed, acyclic graph* $\mathsf{G}' \in \mathsf{P}(\mathsf{G},\mathsf{C})$*, where* G *is an asymmetric structure with a maximal core* C*, has a maximal element in* V^{C}*.*

Let $\mathsf{G}' \in \mathsf{P}(\mathsf{G},\mathsf{C})$. Given that G is a directed, acyclic graph, it has a maximal element m_{G} that is the first element of its core C, by definition. Since G' is a substructure of G, the the transitive closure of $\mathsf{E}_1^{\mathsf{G}}$ must have an edge (m_{G},d) for each $d \in \mathsf{G}'$. By definition of $\mathsf{P}(\mathsf{G},\mathsf{C})$, the edges in G' cannot be in C. Since G' is also a directed, acyclic graph it must have a maximal element m_d. But then any edge $(s,m_d) \in \mathsf{E}_1^{\mathsf{G}}$ is in $\mathsf{E}_1^{\mathsf{C}}$, and we know that there is at least one such edge (m_{G},m_d). Thus, $m_d \in \mathsf{V}^{\mathsf{C}}$. Hence, Fact 1 follows.

Fact 2 *Let* G *be an asymmetric structure with a maximal core* C*. Then* $\mathsf{P}(\mathsf{G},\mathsf{C})$ *contains no edges* $e \in \mathsf{E}_1^{\mathsf{P}}$ *such that* $e = (a,b)$*,* $a \in \mathsf{V}^{\mathsf{P}}$ *and* $b \in \mathsf{V}^{\mathsf{C}}$*.*

The proof of this fact follows immediately from Fact 1 and the observation that were $\mathsf{P}(\mathsf{G},\mathsf{C})$ to contain an edge (a,b) with $a \in V^{\mathsf{P}}$ and $b \in V^{\mathsf{C}}$, then $\mathsf{P}(\mathsf{G},\mathsf{C})$ would contain a chain that could be added to C to obtain a core of which C was a proper substructure. This would contradict the fact that C is maximal.

In principle, a discourse structure can contain several cores; this can happen, for example, in conversations that include multiple threads (Wang et al. 2011). In our corpus, however, the vast majority of conversations have one clearly defined core (given by the game moves and interleaved linguistic moves). In fact, given that the

playing of the game is the primary concern of our dialogue participants (and each has as his or her overall goal to win the game), we conclude that for every discourse graph G for our corpus, *all* of the nonlinguistic game EEUs in G are contained in the core of G. This asymmetry between the nonlinguistic events and linguistic moves might be different for other dialogues. Someone using facial expressions and gestures to react to a speech or story they are hearing, for example, might contribute to an asymmetric structure if these nonlinguistic movements have no effect on the associated linguistic moves.[11] In this case, the relationship between nonlinguistic events and discourse moves would be reversed compared with the *Settlers* corpus: the linguistic moves would by and large form the core and the nonlinguistic ones, the periphery.

5.1.2 New patterns of attachment

The asymmetric structures in our *Settlers* corpus permit discourse attachments that yield "rectangular" discourse structures that we have not encountered in previous annotation campaigns on single-authored texts and which are not countenanced in other coherence-based theories of discourse (indeed, RST is restricted to trees; see Mann & Thompson 1987). This attachment pattern arises when one dangler attaches to another dangler that extends from a separate node of the core. To illustrate this, consider (13), which involves a fragment of a negotiation dialogue that ends with a nonlinguistic move:

(13)

341	Server	GWFS rolled a 6 and a 3.
342	Server	inca gets 2 wheat. dmm gets 1 wheat.
344	GWFS	9 nooo!
344.0.1	UI	GWFS ended their turn.
344.0.2	Server	It's inca's turn to roll the dice.
345	Server	inca rolled a 1 and a 3.
346	Server	CheshireCatGrin gets 1 ore, 1 wood. GWFS gets 2 wood.
347	GWFS	4 better :)
348	Server	inca ended their turn.

(13) yields an asymmetric structure depicted in Figure 4 below. The game advances without interference from the chat moves, which provide only commentary on the game. The core of the structure is the chain of connected units $[341 \rightarrow 342] \rightarrow 344.0.1 \rightarrow 344.0.2 \rightarrow [345 \rightarrow 346] \rightarrow 348$. Intuitively, 344 is a stand-alone comment

11 For a real-life example, watch the GIF embedded in the following article from *The Washington Post*: https://www.washingtonpost.com/news/the-intersect/wp/2017/04/14/5-questions-for-a-washington-post-reporter-whose-eyebrows-became-a-meme/?utm_term=.f44c57d81ddb. The meaning of the reporter's facial expressions remains highly underspecified, but we certainly get a clear sense of her overall reaction to Sean Spicer's comments.

on the dice roll in 341 or the CDU containing the dice roll and the distribution of wheat (which is what makes the roll of a 9 disappointing for GWFS); that is, it is a dangler. Similarly, turn 347 is a comment on 345 and 346, and it is not picked up by subsequent discourse. Unlike a normal dangler, however, it is also intuitively related to the previous dangler, 344. In fact, given the established ways in which discourse structure is used to constrain the interpretation of anaphora and other elided constructions (Hobbs 1985, Polanyi 1985, Asher 1993, Kehler 2002), it must be related to 344 so as to resolve the linguistically implicit arguments of the relation *better* to their intuitive values.

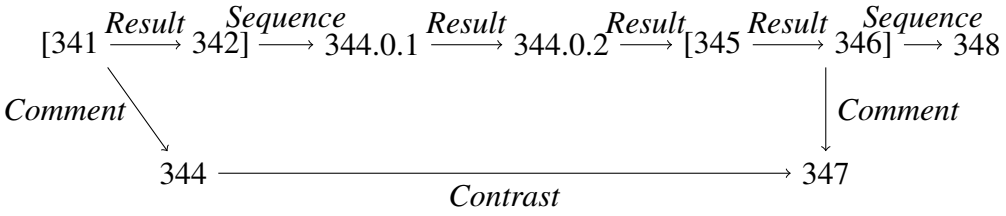

Figure 4: The discourse graph for (13)

Example (14) illustrates the same point, but with a different relation between the danglers and an explicit discourse connective. Note that we have not included the full game sequence of events that form the core of the structure for this dialogue.

(14)

237	Server	dmm rolled a 6 and a 1.
238	GWFS	I can't take another 7.
239	Server	dmm will move the robber.
241.0.1	Server	dmm moved the robber, must choose a victim.
241.0.2	Server	dmm stole a resource from GWFS
244	GWFS	because you keep thieving me

(14) is a semi-constructed example. In the original, GWFS says, "also you keep thieving me" rather than 244. Still, we find (14) to sound natural, and it perhaps more clearly illustrates the rectangular structures we are describing. Turns 238 and 244 are comments on game events, and so dangle off the game moves; at the same time, 244 relates to 238 via Explanation, as indicated by the explicit connective *because*, yielding a rectangular structure.

The preceding examples concern structures with maximal or even unique cores. When a discourse unit d bears multiple discourse relations to different elements $d_1, \ldots d_n$ in a structure G, however, more than one edge (d, d_i) may serve to define a core of G. In such a case, that link is part of one core but perhaps not part of

another, more minimal core. In addition, the structure that includes those elements that involve the main purpose of the dialogue, the playing of the game, may be such a nonmaximal core. Thus, a pair of DUs may support one discourse relation in the periphery and another in the core.

In example (10), repeated below, we have a concrete example of a relation instance *r* between two elements of the core that is itself part of the periphery.

(10)

534	GWFS	anyone want to trade their ore for my wood?
535	LJAY	nope
538	GWFS	it may prove a prudent trade, lj...
539	LJAY	nope
539.1	Server	GWFS played a Soldier card.
539.4	Server	GWFS stole a resource from LJAY
540	GWFS	apologies...
541	LJAY	:(

The failed trade negotiation (turns 534-539) results in GWFS's playing a Soldier card and stealing from LJAY (539.1-539.1). As explained in Section 4.1, the chat moves and the game moves are a part of the game; this is a classic example of an interleaved structure. In addition, however, we infer an Explanation between GWFS's utterance of "it may prove a prudent trade, lj..." (538) and his Soldier card play, and this relation is *not* central to advancing the game play; the *raison d'être* of his move is not to provide an explanation of his warning, so the Explanation is structurally independent of the game structure. The explanatory role played by 539.1 and 539.4 towards 538 could be removed without endangering the coherence of the core, while removing the core from this example would leave us wondering why GWFS asserted 538 and why LJAY is sad in 541.

The graph below gives the representation of (10); the dashed arrow represents the only game-independent edge in this example. To improve the readability of the graph, we note that following SDRT and Polanyi 1985,[12] we distinguish *subordinating* edges, which are labelled with names for subordinating relations such as Elaboration, Explanation, and Background, from *coordinating* edges, which are labelled with coordinating relations such as Contrast, Narration (Sequence),[13] and Result. Coordinating edges are generally represented with horizontal arrows, while subordinating edges are generally represented by vertical arrows. The import of the subordinating/coordinating distinction will be brought out in the next subsection.

12 Rhetorical Structure Theory (Mann & Thompson 1987) also has a similar distinction.

13 In our annotations on the nonlinguistic context, we have opted for Sequence over SDRT's Narration, as the game events do not figure in a narrative in the traditional sense. However, the semantics of Sequence and Narration are exactly the same.

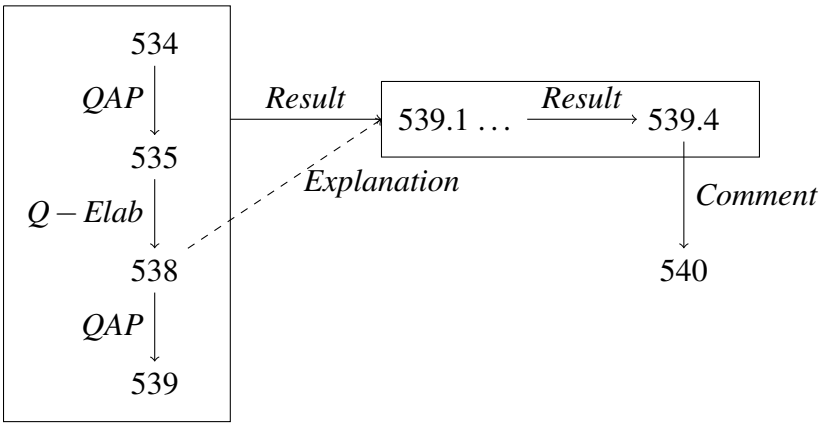

Figure 5: The asymmetric structure for (10). The Explanation edge (538,539.1), represented with a dashed line, is the only game-independent (peripheral) edge.

The asymmetric configuration illustrated in this graph is another example of a structure that is unfamiliar from work on rhetorical structure.

To make the point abstractly, consider a relational structure R with two relations in R represented extensionally as sets of pairs, i.e., $\mathsf{R} = (\{a,b,c\},\{(a,b),(a,c)\},\{(a,b),(b,c)\})$. Removing one of the pairs, (a,b), will yield two substructures R_1 and R_2 of R: $\mathsf{R}_1 = (\{a,b,c\},\{(a,c)\},\{(a,b),(b,c)\})$, and $\mathsf{R}_2 = (\{a,b,c\},\{(a,b),(a,c)\},\{(b,c)\})$. For a given discourse unit u with such multiple relations, one of these rhetorical relations might be a part of the core substructure, while another might contribute to the periphery. That one and the same pair of discourse units can be elements of both the core of an asymmetric structure G and of the periphery of G might seem strange. Nevertheless, the situation is consistent with our core/periphery distinction, which yields a partition of the edges in G but not of its nodes.

5.1.3 Discourse salience and the right frontier

Rectangular structures like that given by (13) violate the Right Frontier Constraint (RFC) for monologue or text (Polanyi 1985, Webber 1988, Asher 1993). The Right Frontier (RF) is a set of nodes in a discourse structure — the nodes along the right edge of a discourse graph — that dynamically evolves as a discourse proceeds, and is designed to track the accessible and salient nodes in a discourse at any given time. The RF *Constraint* requires a new node to attach to a node from the RF.[14]

14 Attachment to nodes that are no longer on the RF is permitted if the jump backwards is explicitly indicated by, for example, repeating content that is no longer on the RF or using a phrase such as, "Let's go back to the second point that you made". Asher 1993 calls this *discourse subordination*.

Normally, a coordinating relation such as Sequence or Result is understood as pushing the discourse forward, thereby shutting off the accessibility of its left argument. To make this more precise, we introduce SDRT's Right Frontier Constraint; RFCs from other accounts are roughly similar. The RF in SDRT includes (i) the node Last — which is the node most recently attached to the discourse graph — as well as (ii) any unit that is superordinate to a unit on the RF through a subordinating relation and (iii) any CDU that includes a node on the RF.[15] Definition 6 formalizes this definition, using the definition of a discourse graph from Definition 1. Definition 6 defines those nodes x on the RF of an SDRT graph G, written $\text{RF}_{\mathsf{G}}(x)$, which are accessible for the next unit to rhetorically attach to:

Definition 6 *Let* $\mathsf{G} = (\mathsf{V}, \mathsf{E}_1, \mathsf{E}_2, \ell, \mathsf{Last})$ *be a discourse graph;* $\forall x \in \mathsf{V}$, $\text{RF}_{\mathsf{G}}(x)$ *iff* ***(i)*** $x = \mathsf{Last}$*; or* ***(ii)*** $\exists y \in \mathsf{V}\ \text{RF}_{\mathsf{G}}(y)$ *and* $\exists e \in \mathsf{E}_1$, $e(x,y)$ *and Sub(e); or* ***(iii)*** $\text{RF}_{\mathsf{G}}(y)$ *and* $\exists e \in \mathsf{E}_2$, $e(x,y)$.

The definition of the RF entails that attaching to the RF via a subordinating (*Sub*) relation places both arguments on the RF, but attaching via a coordinating relation removes the first argument from the RF, leaving it inaccessible for further rhetorical qualification. Many of the rectangular structures in our corpus violate this constraint. The rectangular structure produced by (13), as pictured in Figure 4, is one example: the Result and Sequence relations intervening between 342 and 345 should render 344 inaccessible to future discourse moves, yet the comment in 347 is still interpretable as standing in a Contrast relation with 344.

If we look more closely at the rectangular structures in the *Settlers* corpus, we see that coordinating relations between nonlinguistic events do not necessarily shut off the accessibility of prior nonlinguistic events, either, which gives rise to further violations of Definition 6. Examples like (12), repeated below, are frequent in our corpus. In (12), turn 155 is attached to 154.1-154.3 via Sequence. The attachment of 155 should render 154.1-154.3 inaccessible, yet the Explanation provided by 159 makes clear that 158 should be understood as a Comment on those earlier moves. Likewise, 163 is understood as a Comment on 159.1-159.4 despite the intervening coordinating relations.

15 The RFC is a constraint on accessibility and as such, it generally fails to predict *the* node to which the current unit will attach. The full definition of the constraint on accessibility is further complicated by the presence of the relations *Contrast* and *Parallel*, but we gloss over that here.

(12)

154.1	Server	GWFS played a Soldier card.
154.3	Server	GWFS stole a resource from LJAY
155	Server	GWFS rolled a 5 and a 1.
157	Server	GWFS built a settlement.
158	GWFS	sorry laura
159	GWFS	needed clay the mean way :D
159.1	Server	LJAY played a Soldier card.
159.4	Server	LJAY stole a resource from GWFS
160	Server	LJAY rolled a 4 and a 4.
161	Server	GWFS gets 2 wheat.
163	GWFS	touché

At the same time, while rectangular structures show that the structure of game moves does not effect the development of a chat in the way predicted by Definition 6, we should not conclude that the structure of game moves is entirely irrelevant to chat development. Consider (15), a minimal variant of (12).

(15)

159.1	Server	LJAY played a Soldier card.
159.4	Server	LJAY stole a resource from GWFS
160	Server	LJAY rolled a 4 and a 4.
161	Server	GWFS gets 2 wheat.
162	GWFS	**finally some wheat**
163	GWFS	touché

The introduction of 162 in (15) makes it far more difficult to return to the prior Soldier card event, thereby making 163 less coherent in (15).

The discussion of (12) and (15) brings out a general point about situated discourse. The original RFC, conceived as a constraint on how a speaker should present information and as a constraint on monologue and text, has strong empirical support (Polanyi 1985, Afantenos & Asher 2010). In situated discourse, however, the development of a discourse is not always under the speaker's control. We need to recast constraints on discourse attachment as constraints not only on how information can be *presented*, but on how it can be *exploited*. Already, moving to multi-party dialogue introduces complications for the RFC not only because speakers can engage in multiple, separate threads of conversation (see also the interruption moves from Polanyi 1985), but also because an interlocutor might not agree that information that a speaker has presented as the most salient is the information that should be discussed in the subsequent discourse. Consider the following constructed example:

(16)

100	T.K.	Anyone want ore for sheep?
101	GWFS	**I'm not giving up my sheep for now**.
		lj might want to give some of hers, though.
102	GWFS:	?? Not for all the ore in the world.

Had GWFS only uttered (typed) the words in bold in 101, the move in 102 would have been coherent. Once he utters the words in italics, however, accessibility of the boldface move is shut off, as the two moves are related via Contrast, making 102 highly anomalous. Surprisingly, however, move 102′ in (17) is perfectly felicitous even though it builds directly off the boldface move, ignoring the italicized move.

(17)

100	T.K.	Anyone want ore for sheep?
101	GWFS	**I'm not giving up my sheep for now**.
		lj might want to give some of hers, though.
102′	T.K.	What if I offer you two ore?

This is, we think, because T.K. is not immediately committed to the discourse structure that GWFS builds: he can effectively ignore some aspects of GWFS's prior commitments while addressing others with his own move.

More generally, if an interlocutor has not had a chance to object to a speaker's development of a discourse, he cannot be taken to be committed to the discourse structure that a speaker has laid out. Once the interlocutor utters something that builds off of that structure, however, he indicates his commitment to the structure up through that point. Interactions in our *Settlers* corpus indicate that something similar happens with multimodal interactions: the very fact that one event occurs in a sequence after another event does not mean that it will be more salient for interlocutors observing the events. Once a speaker chooses to appropriate a nonlinguistic event for the purposes of conversation, however, she makes that event salient and thereby commits to a salience ordering over the structure of the events leading up to that salient point.

To extend the RFC to asymmetric structures,[16] we combine this general observation about discourse with our observations about rectangular structures. We propose the following hypothesis: a discourse unit m in the core C of a graph G for an asymmetric structure will remain accessible to a new move n in the periphery of G so long as no commentary is made on m or any move from C that is subsequent to m — this is regardless of whether m corresponds to a linguistically-specified clause (i.e., m is an EDU), a nonlinguistic event (EEU), or an extended unit consisting of (coherently related) sub-units (CDU). In addition, all of the nodes on the RF of the periphery will

16 This extension might be best suited for asymmetric structures with a largely nonlinguistic core. We have not done empirical work on asymmetric structures with linguistic cores, and it may be that while certain violations of the classic RFC are allowed for such structures, they nevertheless impose stricter conversational principles than asymmetric structures with nonlinguistic cores.

remain accessible to n. Once the content of a *linguistic* move n is attached to a node m from C, however, then the RF of the preceding discussion disappears, unless n attaches to it as well, and all nodes in C that should be inaccessible from m according to Definition 6 also become inaccessible to n.

Definition 7 for a *revised right frontier* (RRF) below formalizes these observations.

Definition 7 *Let* $\mathsf{G} = (\mathsf{V}, \mathsf{E}_1, \mathsf{E}_2, \ell, Last)$ *be an asymmetric discourse graph; let* $Acc(\mathsf{G})$ *be the set of labels on the RF of a graph* G *as in Definition 6, and let* $\mathsf{G}^c = (\mathsf{V}^c, \mathsf{E}_1^c, \mathsf{E}_2^c, \ell^c, \mathsf{Last})$ *be the maximal core of G. Then:* $\textsc{rrf}(\mathsf{G}) = Acc(\mathsf{G} - \mathsf{G}^c) \cup \{u \in \mathsf{V}^c \colon \neg\exists e \in (\mathsf{E}_1 \setminus \mathsf{E}_1^c)\ \exists y \in \mathsf{V}^c (y \in End(e) \wedge u \prec y)\}$

This definition captures the idea that in situated conversation, nonlinguistic events constrain coherent discourse progression only when speakers choose to exploit them in the conversation. Interactions between noncommunicative, nonlinguistic events are not in and of themselves controlled by salience constraints; such events generally unfold according to a different set of rules, such as those governed by physical laws or, in the case of our corpus, proper game play. Once speakers appropriate them as a part of their message (by making linguistic moves that coherently connect to them), these events enter into structural relations with other discourse units and speakers are responsible for how they build that structure every bit as much as an author of a newspaper article is responsible for structuring the information that she presents.

In effect, discourse participants make EEUs salient — that is, they determine how these events are brought into the Right Frontier — via their decisions on what they *talk* about. This is clearly not all there is to say about salience: specifically, it ignores how *visual* salience affects accessibility and reference description (see, for instance, Clarke et al. 2015). But we leave this topic for another time.

5.2 The semantics of situated discourse structures

While interleaved structures will be subject to the RFC defined in Definition 7, they highlight a different limitation of extant theories of discourse. Consider (18):

	123	dmm	anybody willing to give me a wood? i can trade clay or ore for it
	124	GWFS	no woods sorry
	126	inca	sorry, none here
	127	LJAY	illl have a clay for one
(18)	127.0.1	dmm	made an offer to trade 4 clay for 1 wood from the bank or a port.
	128	Server	dmm traded 4 clay for 1 wood from the bank.
	129	LJAY	or not
	130.0.1	UI	dmm ended their turn.
	131	dmm	oh well

After receiving two refusals to his offer, dmm trades all of his clay with the bank before waiting for LJAY's response. (We assume that he had started setting up his offer and so missed her reply in the chat window.) Once dmm does this, the trade is no longer a possibility, as he has traded all of his clay away. He regrets his decision, as we see in turn 131, but there is nothing that either he or LJAY can do to repair this situation given that dmm is out of clay and then ends his turn; all they can do is talk about it, which they do in 129 and 131. In other words, once dmm shuts off the possibility of the trade (in move 128), LJAY's offer still remains *salient* in terms of the RFC, but certain types of continuations are shut off on the basis of the game's structure — dmm cannot now accept LJAY's offer because he's out of clay and he's ended his turn.

This observation brings out a semantic point deeper than Hypothesis (NDU). (NDU) is consistent with a modular approach to discourse structure that can represent the coherent connection between a sequence of EEUs and a sequence of EDUs in order to represent the entire semantic content of an interaction; it is, for example, consistent with a case in which interlocutors discuss a problem, formulate a plan, then carry out the plan via nonlinguistic actions and then come back to comment on or discuss the events that took place, and so on. While we have known for a while that discourse structure is generally not isomorphic to an externally given plan structure (Moore & Paris 1993), this modular approach to negotiations is still largely assumed (as, for instance, in classic work on bargaining, such as Osborne & Rubinstein 1990). (18) shows that even this kind of modular approach to interaction is too simple. In our interleaved structures, the modular approach breaks down, because linguistic discussion that would bear on future actions might be unfinished at the point at which a nonlinguistic action eliminates options that the linguistic discussion would have made optimal. It is not simply that we must incorporate the semantic contents of certain nonlinguistic events into semantic structures in order to capture the full content of an interaction; we also need to rethink the update mechanisms involved in interpreting these structures.

In classic dynamic semantics, a context or information state is a set of worlds or variable assignment functions, a set of world-assignment pairs, or a set of such sets (Groenendijk et al. 1996). Update with an assertion changes a given context generally by removing worlds from the incoming set that are incompatible with its content: that is, successive interpretations of contexts in a set C induce a monotone decreasing function f on the world components C_w of those contexts C. In other words, for any $c_w \in C_w$, $f(c_w) \subseteq c_w$.

Of course, in a conversation, interlocutors might hear or interpret things said in a discourse in different ways, or one interlocutor might simply not believe something that another speaker said, and so refuse a proposed update to the context. For this reason, more recent dynamic accounts of discourse interpretation allow each interlocutor in a discourse to have her own, dynamically evolving representation of the discourse. Ginzburg (2012) proposes individual dialogue gameboards for different participants, and SDRT tracks individual commitments using distinct representations (Lascarides & Asher 2009) or dynamic modal operators (Venant & Asher 2016), to give just a few examples.

While these accounts provide a rich and powerful arena for defining how moves in a dialogue affect dynamic information growth, they all in effect assume that the space of possible updates is determined by what can be *said* in a discourse, or by the content of each interlocutor's representation of, and commitments to, what has been said. The actual world plays a very passive role in the various definitions of discourse update to date: if the actual world figures in the set of worlds that survive update with a proposition p, then p is true; if it doesn't, then p is false. But p's being false has no effect on the way the discourse can proceed; all that matters is whether discourse moves are consistent with one another. Speakers are free to say whatever they want, and be wrong. The real world doesn't impinge at all on conversational continuations.

While speakers are free to say things that are false, when the contents of events that are actually taking place in the world start to interact with discourse moves, the actual world begins to play a much more active role in limiting possible continuations. The actual world cannot be inaccurate, nor can it be inconsistent. Once dmm gives his clay to the bank in (18), he cannot give any clay to LJAY later on, unless he first gets more clay.[17] This affects possible continuations of speech acts. Before dmm trades his clay away, he could have responded to LJAY's answer with

17 In fact, it's not even true that the set of possible continuations is entirely open for linguistic contexts. Once you have said something, you cannot *unsay* it. You can claim or even believe that you never said it, but there is no continuation of the actual linguistic context in which you never made that commitment. Speech acts change the truth about what was said, but they, like nonlinguistic events, also change the world and potentially the truth of one's first order commitments. This fact about speech acts, however, is not usually of central concern in the development of discourse models.

an offer to trade and she could have accepted the trade, which would have left dmm in a strategically preferable position. Once dmm trades his clay away, a continuation of the game in which he gives his clay to her is no longer possible. Both dmm and LJAY are free to talk about the offer and build linguistic continuations off it, but this discussion is inert with regard to game development. In other words, moves building on LJAYS's response can only contribute to the periphery of an asymmetric structure. The world is not there simply for interlocutors to reflect on and learn about; if we are too slow in our discussions, we are liable to find that the world has moved on without us, and we will need to readjust the set of possibilities.

Once the contents of nonlinguistic actions are a part of the semantic representations of multimodal interactions, as we have proposed they should be in this paper, then these limitations on update need to be made explicit. Dynamic update must involve not only an evolution of the set of possible worlds, but also a dynamic evolution of the worlds themselves. For our conversationalists, the world changes as time goes on, in part due to their actions, in part due to physical processes.[18] The actual world allows for certain possible futures, in virtue of which we decide to act. But once we act, some of those possibilities become closed off. To make this concrete, we introduce three new ingredients into our dynamic semantic models. First, we add a function $\mathfrak{h}$ that maps each world w to a *history*, where a history is a finite sequence of events that occur in w. Second, we add a set of rules L that constrains how histories can develop; in a model of *Settlers* for instance, L determines the legal game sequences. Finally, we need to link the events denoted by DU labels with the contents associated with them in logical form. For this we add a relation S_w that links the denotation of each DU or EU to a semantic content in a world w. Because EDUs and CDUs formed from them express propositions that are not about the speech acts these DUs themselves denote, we cannot make use of Davidson's (1967) proposal about action sentences to handle this linking. On the other hand, EEUs are different: the formula that provides their content *does* characterize them, and so we can make use of Davidson's proposal for EEUs. That is, where ε is an EEU, a formula of the form '$\varepsilon\colon\ \phi$' means that ϕ characterizes the eventuality itself (as opposed to characterizing content conveyed by a speech act).

More precisely, we assume a linear temporal order $\leq_t$ over EEUs ε derived from their ordering in situated conversation; thus $\varepsilon_n \leq_t \varepsilon_m$ for $n < m$ (cf. Section 5.1.3). A model $\mathfrak{A}$ for a graph G representing a conversation with n players will then be a tuple $\mathfrak{A} = \langle D_i, D_e, W, C_1, \ldots, C_n, S, \mathsf{L}, \mathfrak{h}\rangle$, where D_i is the domain of individuals including a set of players, D_e is the domain of events, W is the set of worlds, and for each player i, C_i is an accessibility relation. $\mathfrak{h}\colon W \to (\mathscr{P}(D_e))^*$ is a function giving the "history"

18 Note that this makes the actual world, indeed all worlds, a kind of branching structure; for the *Settlers* corpus the actual world is the game tree plus possible conversational events, which gets whittled down as the players play.

of a world, where $(\mathscr{P}(D))^*$ is the set of all finite sequences of sets of eventualities. $\mathfrak{h}_m(w)$ is the history of w restricted to the first m moments. Assignment functions will map individual variables into D_i and DU variables into D_e.

Update with the content of an EEU will proceed in contexts that consist of an assignment function f and a world w, where w determines a set of commitment slates $C_i(w)$ for each player i and a history $\mathfrak{h}(w)$. Let a formula of the form '$\varepsilon\colon\ \phi$' mean that ε is characterized by the content ϕ. Then an EEU $\varepsilon\colon\ \phi$, where ε occurs at moment n (written ε_n) will update the context by minimally: (i) extending the assignment function f to an assignment f' whose domain is that of f plus ε_n; (ii) shifting the world of evaluation w to a world w' such that (a) w' complies with L and $f'(\varepsilon_n) \in \mathfrak{h}_n(w')$ ($f'(\varepsilon_n)$ is included in the set of eventualities at n) and $\mathfrak{h}_m(w) = \mathfrak{h}_m(w'), \forall m < n$ (i.e., the past history remains unchanged), and (b) (w', f') verifies ϕ. This means that while worlds in the update may differ on commitments and perhaps even what events they contain, they must all contain the events introduced by DUs in the order in which they were introduced. Our procedure guarantees that the actual world remains in the context set upon EEU update.

Update with EDUs works analogously; an update with an EDU or EEU transforms the world — the world has an event in it that it did not have before. However, EDUs and EEUs differ in an important respect: unlike EEUs, a world w with the appropriate history and assignment f need not satisfy the content ϕ associated with an EDU π in the SDRT formula $\pi\colon\ \phi$ (i.e., its context change potential lacks conjunct (b) in clause (ii) above). Interlocutors are still free to say and to commit to contents that are false. Details are in the Appendix.

As situated conversation is still conversation, we need to also say something about how commitments evolve in our model. A player i's commitments at a world w change when updated with an SDRT formula $\pi^i\colon\ \phi$. We take a minimal view to commitment change made by discourse actions here.[19] We assume that i at least publicly commits to the conventional meaning of her verbal message and to the fact that her speech act π^i has the content assigned to it by the semantics. In addition, an interpreter j who exploits a discourse move π by speaker i to link her own contribution to the conversation will commit that i commits to the content of π; in particular, if j links i's contribution π with a relation R to some other DU ρ, j will commit to $R(\rho, \pi)$ but also to the content that i commits to π. i may interpret matters differently and claim she was committed to something different. Her SDRS for the conversation would then differ from j's (Lascarides & Asher 2009).

With regard to EEUs, no speaker need commit to the basic content ϕ of an EEU $\varepsilon\colon\ \phi$, unless: (i) she is causally responsible for ε, or (ii) she makes ε part of the discourse structure by relating it to another DU ρ via a discourse relation R. In the

19 For a full treatment of nested, higher-order commitments, see Venant & Asher 2016.

latter case, i commits to the content of $R(\varepsilon, \rho)$, which in turn might commit her to the content associated with ε, depending on whether R is a relation that entails the dynamic conjunction of the contents of the units it connects, or not. Most moves that involve relations to EEUs, like Result, Explanation and QAP, are *veridical*, which means they do entail the dynamic conjunction of the contents of their arguments.

To illustrate our semantics, let's return to (18).

(18)

123	dmm	anybody willing to give me a wood? i can trade clay or ore for it
124	GWFS	no woods sorry
126	inca	sorry, none here
127	LJAY	illl have a clay for one
127.0.1	dmm	made an offer to trade 4 clay for 1 wood from the bank or a port.
128	Server	dmm traded 4 clay for 1 wood from the bank.
129	LJAY	or not
130.0.1	UI	dmm ended their turn.
131	dmm	oh well

Turn 123 introduces a question that commits dmm to two possible (sets of) continuations, one in which someone gives him a wood and one in which no one does. The second sentence of 123 introduces an elaboration on the exchange dmm envisions. Turns 124 and 126 commit GWFS and inca to the fact that dmm has so committed and they also commit to not offering him a wood. In turn 127, on the other hand, LJAY commits to a continuation in which she does the exchange with dmm. In our semantics, the actual world is still compatible with this exchange happening, in the sense that the actual world is an element of the continuation in which the exchange takes place. However, in turns 127.0.1 and 128, dmm sets up and completes a trade with the bank. Our semantics predicts that the world now changes or shifts, and some possible continuations in which the actual world figured prior to the exchange with the bank are no longer possible. In particular, dmm has given away all his clay and so he cannot trade with LJAY even though he intended to trade with someone and LJAY was willing.

The discourse structure partially models this, since it features the relation Result(π,127.0.1), where π is a CDU consisting of the (coherently related) segments 123-126. The semantics of this relation entails that π's two negative responses to the trade offer cause the nonlinguistic action described in 127.0.1 of dmm trading with the bank. It also implies that dmm commits to the negative responses by GWFS and inca as well as to the result relation between this (failed) trade negotiation and her bank trade. But this also means that dmm *does not commit* to LJAY's response in 127. We note that given the way the semantics is set up in the Appendix, nei-

ther GWFS nor inca need commit to the result or the bank trade offer, as intuitions dictate (they might not have been paying attention). In 129, LJAY commits to the new real-world event of the bank trade by commenting on it. Finally, in 131, dmm now realizes his mistake; by commenting on LJAY's turn in 127, he commits to her positive response to his offer.

6 Related work on discourse structure

Much of this paper has been dedicated to a discussion of how semantic interactions with nonlinguistic events can give rise to new kinds of semantic structures with their own constraints on discourse evolution and interpretation. It complements work on multi-party dialogue that has compared features of multi-party dialogue with monologue (Ginzburg & Fernández 2005) and that has explored the behaviour of conversational threads (Elsner & Charniak 2011). It also complements work by Goffman (1981), who noted that there may be participants in a conversation d that are not "ratified" by the active participants in d.[20] These participants may eavesdrop on d and then exploit elements of d in their own conversation; for their conversation, moves in d are discourse independent, as we discussed earlier. Our work extends and complements this earlier work by making explicit different structural possibilities that arise in situated discourse, and by investigating the consequences of such interwoven discourses for the study of discourse more generally.

Our work also extends research by Lascarides & Stone (2009) on the rhetorical structure of conversation with co-speech gesture and by Stojnic et al. (2013) on descriptions of unfolding events. Nonlinguistic events are less constrained in their possible rhetorical roles than co-speech gestures; Lascarides & Stone argue, for instance, that one cannot coherently use Contrast to connect a gesture to its coverbal speech, while our *Settlers* corpus has many instances of nonlinguistic discourse units participating in Contrast relations, of which (19) is one example:

(19) Server: player i made an offer to trade 1 wheat for 1 sheep from j
player j: But I don't have any sheep.

Stojnic et al. (2013) focus their analysis on a particular kind of coherence relation between linguistic and nonlinguistic moves: a relationship they call *Summary*, in which the linguistic move describes what is currently happening in the (visual) embodied environment. This specific relationship between language and vision also underpins existing multimodal parsing technologies trained on videos and captions (Yu et al. 2015). In addition, there are systems supporting embodied human robot interaction which use a combination of language and vision to recognise

20 For further discussion see Dynel 2010.

the current state and the user's intentions, which in turn influences the robot's decisions about which actions to perform (Foster & Petrick 2014, Forbes et al. 2015, Liang 2005). These systems effectively combine a natural language instruction with evidence from the visual scene to help specify the specific robot motions that the user requires, and a major part of this process involves grounding the natural language symbols to visually salient entities. Our corpus and examples like *Scratches* and *Table* illustrate that nonlinguistic events enter into a wider range of coherence relations than this. Further, this prior work on video captions and HRI focusses on single isolated utterances and their relationship to the visual scene, and so it has largely bypassed the need to study how the discourse structure of a prior *extended* multimodal conversation, of the kind that the *Settlers* corpus exhibits, constrains successive coherent dialogue moves. These two dimensions to multimodal meaning have been the main focus of our paper: we have explored in detail how incorporating a wide range of coherence relations into the structure of an extended multimodal conversation calls for revisions as to what structures are possible and the model theory for interpreting them.

An alternative to using SDRT to spell out the details of our analysis might be to use a Question Under Discussion (QUD) model, such as those proposed in Ginzburg 2012 or Roberts 2012. It is worth noting, however, that some motivations for adopting a QUD model do not apply to our data. For example, exploiting question-answer congruence to constrain focus (Halliday 1967, Roberts 2012) is not relevant when the contents associated with nonlinguistic events do not have a focus structure in any obvious way. There is also an important lack of parallel between the way in which the kind of EEUs that we have explored become integrated in a discourse structure and the means by which QUDs are posited to contribute to discourse. Explicit QUDs are uncontroversially linguistic, but one should also distinguish implicit QUDs from nonlinguistic events. Firstly, most nonlinguistic events cannot naturally be construed as questions, implicit or otherwise. Secondly, implicit QUDs are semantic contents inferred on the basis of information structural properties of explicit utterances or by considering the relation between two explicit utterances (van Kuppevelt 1996) — they are not the contents of *events* of any sort, and so cannot be expected to have the same effects on update as the nonlinguistic events with which we have been concerned in this paper. The latter cannot in general be inferred based on the surrounding linguistic moves. In *Scratches*, for instance, Anne's utterances no doubt influence Peter's conceptualization of the scratch on the wall, but without the perception of the scratch itself, no amount of reasoning will help Peter figure out why Anne punished their daughter.

These differences aside, QUD, like SDRT, has served to analyze a variety of anaphoric and elided expressions featured in conversation, including sentence fragments (Ginzburg & Sag 2001, Ginzburg 2012). Many of the contributions linked

by Comment in our corpus are sentence fragments or incomplete utterances. Like us, Ginzburg posits that incomplete utterances have as a part of their semantics an anaphoric dependency, and for QUD models this is resolved by linking the fragment to questions that are accommodated as the discourse proceeds, and that determine what the discourse is about.

While to our knowledge there is no QUD-based analysis of the semantic contribution of nonlinguistic events to discourse, given the parallels that can be drawn between Ginzburg's treatment of sentence fragments in QUD and Schlangen's (2003) treatment of fragments in SDRT, one might be able to reconstruct the SDRT account presented here within the QUD framework. One intuitive and useful contribution of QUD needed for analyzing situated communication is that linguistic moves can generate expectations that guide conceptualization. And in *Scratches*, for example, Peter's expectation of an Explanation arguably helps guide him to adopt a certain conceptualization of the explanandum. This "top down" information flow, from expectations about discourse structure to a pragmatic interpretation of an individual unit, contrasts with a common "bottom up" approach where reasoning flows from the observed signal to discourse structure (e.g., Hobbs et al. 1993), though in computational models for SDRT or RST (Muller et al. 2012, Joty et al. 2015), the construction of a discourse representation is modelled as a constraint satisfaction problem and so information flows both bottom up and top down.

Nevertheless, differences between SDRT and QUD analyses of sentence fragments make such reconstruction challenging. Ginzburg's QUD model assumes that some information about a speech act that is performed gets encoded within the linguistic grammar; Ginzburg uses this aspect of the sentence fragment's syntax to constrain its interactions with context, in particular with the *linguistic form* of the context. In contrast, SDRT's approach makes no such assumptions about the linguistic grammar, and the semantics/pragmatics interface has no access to linguistic form, but only to a partial description of the content *derived* from linguistic form. As in *Scratches* and *Table*, the *form* of the nonlinguistic event e that becomes a part of the message may be unobservable to both the speaker and the interlocutor. In such cases, which are common, there is no motivation to make its form a necessary premise to computing its semantic role in the conversation.

7 Conclusions

We have provided empirical evidence that nonlinguistic events participate in conveying a coherent overall message in situated conversation by contributing the contents of entire discourse units to the content of a discourse. We have further argued that coherence-based frameworks of discourse interpretation are ideally suited for modelling these kinds of contributions of nonlinguistic events, and we've laid out the key

steps in extending a coherence-based theory accordingly. Linguistic and nonlinguistic moves are interpreted jointly within an integrated architecture, linking linguistic form and context to meaning (formal details can be found in the Appendix.)

The *Settlers* corpus provides real data to support our rhetorical model. This data is helpful because it manifests a wide variety of rhetorical interactions and a mixture of structural configurations, including asymmetric and interleaved configurations. Moreover, because of the way the corpus was constructed, it provides a consistent context against which we can explore the nature of situated discourse over multiple discourse moves in an extended coherent conversation. While it may be possible to construct such examples, doing so would require also describing the context relevant for each example. The background context provided by the corpus saves us from having to perform this task. In addition, the corpus setup has allowed us to circumvent the conceptualization problem — the question of how linguistic moves influence the conceptualization of nonlinguistic events — which has facilitated the study of the codependent task of determining how conceptualizations of nonlinguistic events affect discourse coherence. In particular, we have been able to examine how nonlinguistic moves affect the salience of other linguistic and nonlinguistic moves and influence the constraints on the hierarchical development of discourse, and the data have shown that these constraints differ from those that apply in linguistically-specified discourse. We hope to study the complex problems of individuation and conceptualization in future work with different data, leading to computational models of situated dialogue parsing.

In future work, we also hope to get a better understanding of how to classify different kinds of nonlinguistic events in order to articulate the different effects those event types have on discourse structure. We briefly noted some ways in which the game events from the *Settlers* corpus differ from coverbal gestures, but for the most part we have treated nonlinguistic events as a homogenous group. This is an idealization. For coverbal iconic gestures, a rudimentary and underspecified conceptualization comes from conventions about the form of the gesture, the form of the speech, and their relative timing (Kendon 1983, Lücking 2016, Alahverdzhieva & Lascarides 2010). For purely nonlinguistic events, the interpreter must retrieve the conceptualization from her visual observations together with the discursive links that speakers provide to these events. Thus, any explicit procedure for building a situated discourse structure with nonlinguistic eventualities, which is what is needed to complete our analysis, would have to involve a perceptual module that can individuate, and offer conceptualizations of, nonlinguistic events and states, following Liang (2005), Larsson (2013), and others. Once nonlinguistic moves have become conceptualized, segmented units that figure in the same discourse relations as linguistically specified units, we think it will be relatively straightforward to

extend prior statistical models estimating the discourse structure of purely linguistic units (Muller et al. 2012) to models that estimate situated discourse structure.

References

Afantenos, Stergos & Nicholas Asher. 2010. Testing SDRT's right frontier. In *The 23rd international conference on computational linguistics (COLING)*, 1–9. Beijing. http://www.aclweb.org/anthology/C10-1001.

Afantenos, Stergos, Nicholas Asher, Farah Benamara, Anaïs Cadilhac, Cedric Dégremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, Philippe Muller, Soumya Paul, Vincent Popescu, Verena Rieser & Laure Vieu. 2012. Modelling strategic conversation: Model, annotation design and corpus. In *The 16th workshop on the semantics and pragmatics of dialogue (SeineDial)*, 145–146. Paris. https://hal.archives-ouvertes.fr/hal-01138035/document.

Alahverdzhieva, Katya & Alex Lascarides. 2010. Analysing language and co-verbal gesture in constraint-based grammars. In *The 17th international conference on head-driven phrase structure grammar (HPSG)*, 5–25. Paris. http://web.stanford.edu/group/cslipublications/cslipublications/HPSG/2010/alahverdzhieva-lascarides.pdf.

Asher, Nicholas. 1993. *Reference to abstract objects in discourse*. Dordrecht: Kluwer Academic Publishers. http://dx.doi.org/10.1007/978-94-011-1715-9.

Asher, Nicholas, Julie Hunter, Mathieu Morey, Farah Benamara & Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: The STAC corpus. In *The tenth international conference on language resources and evaluation (LREC 2016)*, 2721–2727. Portorož: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/summaries/339.html.

Asher, Nicholas & Alex Lascarides. 2003. *Logics of conversation*. New York, NY: Cambridge University Press.

Asher, Nicholas, Soumya Paul & Antoine Venant. 2017. Message exchange games in strategic conversations. *Journal of Philosophical Logic* 46(4). 355–404. http://dx.doi.org/10.1007/s10992-016-9402-1.

Baroni, Marco. 2016. Grounding distributional semantics in the visual world. *Language and Linguistics Compass* 10(1). 3–13. http://dx.doi.org/10.1111/lnc3.12170.

Chambers, Nathanial, James Allen, Lucian Galescu & Hyuckchul Jung. 2005. A dialogue-based approach to multi-robot team control. In *The 3rd international multi-robot systems workshop*, 257–262. Washington, D.C. http://dx.doi.org/10.1007/1-4020-3389-3_21.

Clarke, Alasdair, Micha Eisner & Hannah Rohde. 2015. Giving good directions: Order of mention reflects visual salience. *Frontiers in Psychology* 6(1793). 1–10. http://dx.doi.org/10.3389/fpsyg.2015.01793.

Davidson, Donald. 1967. The logical form of action sentences. In Nicholas Rescher (ed.), *The logic of decision and action*, 81–95. Pittsburgh, PA: University of Pittsburgh Press.

Dobnik, Simon, Robin Cooper & Staffan Larsson. 2013. Modelling language, action, and perception in type theory with records. In Denys Duchier & Yannick Parmentier (eds.), *Constraint solving and language processing*, vol. 8114 Lecture Notes in Computer Science, 70–91. Berlin & Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-642-41578-4_5.

Dynel, Marta. 2010. Not hearing things, AI hearer/listener categories in polylogues. In *mediazioni 9*, http://www.mediazioni.sitlec.unibo.it/images/stories/PDF_folder/document-pdf/2010/dynel_2010.pdf.

Elbourne, Paul D. 2005. *Situations and individuals*, vol. 90. Cambridge, MA: Massachusetts Institute of Technology Press.

Elsner, Micha & Eugene Charniak. 2011. Disentangling chat with local coherence models. In *The 49th annual meeting of the Association for Computational Linguistics (ACL)*, 1179–1189. Portland, OR. https://pdfs.semanticscholar.org/f854/10ae4a64c6acf281f797b1a21337ce3b37ec.pdf.

Forbes, Maxwell, Rajesh Rao, Luke Zettlemoyer & Maia Cakmak. 2015. Robot programming by demonstration with situated spatial language understanding. In *The 2015 IEEE international conference on robotics and automation (ICRA)*, 2014–2020. Seattle, WA. http://dx.doi.org/10.1109/ICRA.2015.7139462.

Foster, Mary Ellen & Ronald P. A. Petrick. 2014. Planning for social interaction with sensor uncertainty. In *The ICAPS 2014 scheduling and planning applications workshop (SPARK)*, 19–20. Portsmouth, NH. https://researchportal.hw.ac.uk/en/publications/planning-for-social-interaction-with-sensor-uncertainty.

Ginzburg, Jonathan. 2012. *The interactive stance: Meaning for conversation*. Oxford: Oxford University Press.

Ginzburg, Jonathan & Raquel Fernández. 2005. Scaling up from dialogue to multilogue: Some principles and benchmarks. In *The 43rd annual meeting of the association for computational linguistics*, 231–238. Ann Arbor, MI: Association for Computational Linguistics. http://www.aclweb.org/anthology/P05-1029.

Ginzburg, Jonathan & Ivan A. Sag. 2001. *Interrogative investigations: The form, meaning and use of English interrogatives*. Palo Alto, CA: Center for the Study of Language and Information (CSLI). https://web.stanford.edu/group/cslipublications/cslipublications/pdf/1575862786.pdf.

Goffman, Erving. 1981. *Forms of talk*. Philadelphia, PA: University of Pennsylvania Press.

Groenendijk, Jeroen. 2003. Questions and answers: Semantics and logic. In *The 2nd CologNET-ElSNET Symposium. Questions and Answers: Theoretical and Applied Perspectives*, 16–23. Utrecht. https://pure.uva.nl/ws/files/3604214/27248_groenendijkquestanansw.pdf.

Groenendijk, Joroen, Martin Stokhof & Frank Veltman. 1996. This might be it. In Jerry Seligman & Dag Westerståhl (eds.), *Language, logic and computation: The 1994 Moraga proceedings*, 255–270. Palo Alto, CA: Center for the Study of Language and Information (CSLI).

Halliday, Michael A.K. 1967. Notes on transitivity and theme in English: Part 2. *Journal of Linguistics* 3(2). 199–244. http://dx.doi.org/10.1017/S0022226700016613.

Hobbs, Jerry R. 1979. Coherence and coreference. *Cognitive Science* 3(1). 67–90. http://dx.doi.org/10.1207/s15516709cog0301_4.

Hobbs, Jerry R. 1985. On the coherence and structure of discourse. Tech. rep. Center for the Study of Language and Information (CSLI) Palo Alto, CA. https://www.isi.edu/~hobbs/ocsd.pdf.

Hobbs, Jerry R., Martin Stickel, Douglas Appelt & Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence* 63(1–2). 69–142. http://dx.doi.org/10.1016/0004-3702(93)90015-4.

Hunter, Julie. 2014. Structured contexts and anaphoric dependencies. *Philosophical Studies* 168(1). 35–58. http://dx.doi.org/10.1007/s11098-013-0209-4.

Hunter, Julie, Nicholas Asher & Alex Lascarides. 2015. Integrating non-linguistic events into discourse structure. In *The 11th international conference on computational semantics (IWCS)*, 184–194. London. http://www.aclweb.org/anthology/W15-0123.

Joty, Shafiq, Giuseppe Carenini & Raymond Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics* 41(3). 385–435. http://dx.doi.org/10.1162/COLI_a_00226.

Kamp, Hans. 1981. The paradox of the heap. In Uwe Mönnich (ed.), *Aspects of philosophical logic*, 225–277. Dordrecht: Springer. http://dx.doi.org/10.1007/978-94-009-8384-7.

Kamp, Hans. 1990. Prologomena to a structural theory of belief and other attitudes. In C. Anthony Anderson & Joseph Owens (eds.), *Propositional attitudes: The role of content in logic, language and mind*, 27–91. Palo Alto, CA: Center for the Study of Language and Information (CSLI).

Kaplan, David. 1989. Demonstratives. In Joseph Almog, John Perry & Howard Wettstein (eds.), *Themes from Kaplan*, 481–563. Oxford & New York, NY: Oxford University Press.

Kehler, Andrew. 2002. *Coherence, reference and the theory of grammar*. Palo Alto, CA: Center for the Study of Language and Information (CSLI).

Kendon, Adam. 1983. Gesture and speech: How they interact. In John Wiemann & Randall Harrison (eds.), *Nonverbal interaction*, 13–46. Beverly Hills: Sage Publications.

Kranstedt, Alfred, Andy Lüking, Thies Pfeiffer, Hannes Rieser & Ipke Wachsmith. 2006. Deixis: How to determine demonstrated objects using a pointing cone. In *Gesture in human-computer interaction and simulation*, vol. 3881 GW 2005. Lecture Notes in Computer Science, 300–311. Berlin & Heidelberg: Springer.

Larsson, Staffan. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation* 25(2). 335–369. http://dx.doi.org/10.1093/logcom/ext059.

Lascarides, Alex & Nicholas Asher. 2009. Agreement, disputes and commitment in dialogue. *Journal of Semantics* 26(2). 109–158. http://dx.doi.org/10.1093/jos/ffn013.

Lascarides, Alex & Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics* 26(4). 393–449. http://dx.doi.org/10.1093/jos/ffp004.

Lepore, Ernest & Matthew Stone. 2015. *Imagination and convention: Distinguishing grammar and inference in language*. Oxford: Oxford University Press.

Liang, Percy. 2005. *Semi-supervised learning for natural language*. Cambridge, MA: Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology dissertation.

Lücking, Andy. 2016. Modeling co-verbal gesture perception in type theory with records. In *The 2016 federated conference on computer science and information systems*, 383–392. Gdansk: Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.15439/2016F83.

Maier, Emar. 2009. Proper names and indexicals trigger rigid presuppositions. *Journal of Semantics* 26(3). 253–315. http://dx.doi.org/10.1093/jos/ffp006.

Mann, William C. & Sandra A. Thompson. 1987. Rhetorical structure theory: A framework for the analysis of texts. *International Pragmatics Association Papers in Pragmatics* 1. 79–105. https://journals.linguisticsociety.org/elanguage/pip/article/download/144/144-431-1-PB.pdf.

Matuszek, Cynthia, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo & Dieter Fox. 2012. A Joint Model of Language and Perception for Grounded Attribute Learning. In *The 2012 international conference on machine learning*, Edinburgh. https://arxiv.org/abs/1206.6423.

Moore, Johanna D. & Cécile L. Paris. 1993. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics* 19(4). 651–695. https://dl.acm.org/citation.cfm?id=972504.

Muller, Philippe, Stergos Afantenos, Pascal Denis & Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *The international conference in computational linguistics (COLING)*, 1883–1900. Mumbai: The COL-

ING 2012 Organizing Committee. http://www.aclweb.org/anthology/C12-1115.

Osborne, Martin & Ariel Rubinstein. 1990. *Bargaining and markets*. San Diego, CA: Academic Press. http://www.uib.cat/depart/deeweb/pdi/hdeelbm0/arxius_decisions_and_games/bargainingandmarkets.pdf.

Perzanowski, Dennis, Alan Schultz, William Adams, Elaine Marsh & Magda Bugajska. 2001. Building a multimodal human-robot interface. *Intelligent Systems* 16(1). 16–21. http://dx.doi.org/10.1109/MIS.2001.1183338.

Polanyi, Livia. 1985. A theory of discourse structure and discourse coherence. In William Eilfort, Paul Kroeber & Karen Peterson (eds.), *Papers from the general session at the 21st regional meeting of the Chicago Linguistics Society*, vol. 21, 306–322. Chicago, IL: Chicago Linguistics Society.

Roberts, Craige. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5(6). 1–69. http://dx.doi.org/10.3765/sp.5.6.

Schlangen, David. 2003. *A coherence-based approach to the interpretation of non-sentential utterances in dialogue*: University of Edinburgh dissertation.

Stanley, Jason & Zoltan Szabó. 2000. On quantifier domain restriction. *Mind and Language* 15(2-3). 219–261. http://dx.doi.org/10.1111/1468-0017.00130.

Stojnic, Una, Matthew Stone & Ernie Lepore. 2013. Deixis (even without pointing). *Philosophical Perspectives* 27(1). 502–525. http://dx.doi.org/10.1111/phpe.12033.

van Kuppevelt, Jan. 1996. Inferring from topics: Scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy* 19(4). 393–443. https://link.springer.com/content/pdf/10.1007%2FBF00630897.pdf.

Venant, Antoine & Nicholas Asher. 2016. Ok or not ok? Commitments, acknowledgments and corrections. In *Semantics and linguistic theory (SALT 25)*, 595–614. Stanford: Linguistic Society of America and Cornell Linguistics Circle. http://dx.doi.org/10.3765/salt.v25i0.3072.

Venant, Antoine, Nicholas Asher, Philippe Muller, Pascal Denis & Stergos Afantenos. 2013. Expressivity and comparison of models of discourse structure. In *Special interest group on discourse and dialogue (SIGdial 2013)*, 2–11. Metz. http://www.aclweb.org/anthology/W13-4002.

Wang, Li, Marco Lui, Su Nam Kim, Joakim Nivre & Timothy Baldwin. 2011. Predicting thread discourse structure over technical web forums. In *The conference on empirical methods in natural language processing* (EMNLP 2011), 13–25. Stroudsburg, PA: Association for Computational Linguistics. http://www.aclweb.org/anthology/D11-1002.

Webber, Bonnie. 1988. Tense as discourse anaphor. *Computational Linguistics* 14(2). 61–73. http://www.aclweb.org/anthology/J88-2006.

Yu, Haonan, Siddharth Narayanaswarmy, Andre Barbu & Jeff Siskind. 2015. A compositional framework for grounding language inference, generation, and acquisition in video. *Journal of Artificial Intelligence Research (JAIR)* 52(1). 601–713. http://dx.doi.org/10.1613/jair.4556.

Zarrieß, Sina & David Schlangen. 2017. Obtaining referential word meanings from visual and distributional information: Experiments on object naming. In *The 55th annual meeting of the Association for Computational Linguistics (ACL)*, 243–254. http://dx.doi.org/10.18653/v1/P17-1023.

Appendix: Syntax and semantics of situated SDRSs

We extend here the syntax and semantics of the classic SDRT formal language (Asher & Lascarides 2003) to interpret situated discourse structures and show how the definitions apply in the case of the example *Scratches* introduced in Section 1.

The classic SDRT language L_{sdrt} builds on a language L of dynamic semantics with a first order syntax extended with event and individual terms, modalities, and λ-abstractions needed for expressing questions, imperatives, deontic expressions, and attitudes. L_{sdrt} includes a countable set $\pi_1, \pi_2, \ldots$ of *labels* for discourse units and binary rhetorical relation symbols R_n that take these labels as arguments. For situated SDRT, we add to L_{sdrt} a countable set $\varepsilon_1, \varepsilon_2, \ldots$ of labels for EEUs. Each label π and ε is indexed with the agent i who is responsible for that move: for an elementary discourse unit (EDU), the person responsible is the speaker/author; for an elementary event unit (EEU), the person responsible is the one who committed to that action (where EEUs that lack a volitional agent have no superscript); and the person responsible for a complex discourse unit (CDU), which consists of one or more EDUs and EEUs, is the person responsible for its last EDU.

Situated SDRT formulas are defined recursively in terms of L:

Definition 8 SDRT Formulas:

i. *Where ϕ is a formula of L, π: ϕ and ε: ϕ are SDRT formulas;*

ii. *Where ρ_1, ρ_2 are labels and R is a rhetorical relation symbol, $R(\rho_1, \rho_2)$ and $\neg R(\rho_1, \rho_2)$ are SDRT formulas;*

iii. *Where ϕ, ψ are SDRT formulas, $\phi \wedge \psi$ is an SDRT formula;*

iv. *Where ϕ is a conjunction of SDRT formulas, then π: ϕ is an SDRT formula.*

A label π or ε is treated as a discourse referent or existentially bound variable that denotes a speech act or nonlinguistic event, respectively. As explained in Section 5.2, a formula of the form π: ϕ or ε: ϕ is interpreted by a function S_w that maps the denotation of each π or ε to the content of a formula ϕ that serves as the semantic contribution of the unit π or ε at w.

Where i is any speaker, (20)-(25) provide an interpretation of discourse and event units (which may be simple or complex) and their associated formulas.[21] In particular, (21) and (22) formalize the context change potential (CCP) for discourse units and event units, respectively. $f \subseteq_\rho f'$ indicates that f' extends the assignment

21 We forego an analysis of questions here, but note that we could lift the basic semantics to sets of world-assignment pairs: a question would partition the input set of worlds so that each equivalence class in the output partition would correspond to a possible answer (Groenendijk 2003).

f over ρ, where labels $\rho, \rho_1, \rho_2, \ldots$ range over EEUs and EDUs. When an action x is performed in a world w at a time n, the updated world, w_x, will share its history $\mathfrak{h}$ with w up to time n, but will in addition include the speech act(s) or nonlinguistic event(s) denoted by x; that is, $\mathfrak{h}_m(w_x) = \mathfrak{h}_m(w)$ for all $m < n$, and $f'(x) \in \mathfrak{h}_n(w_x)$.

To model speaker commitments, we adopt a modal accessibility relation over world assignment pairs: an agent i's commitment slate, C_i, at a world w relative to an assignment f is the set of all world-assignment pairs accessible from (w, f) via the accessibility relation for i. This allows us to use standard first-order dynamic semantics to define how the content of an EEU or EDU updates a player's base level commitments. Note that when the world of evaluation shifts in (21), it shifts in i's commitment slate C_i as well. Update is defined with relational composition $\circ$.

(20) for $\phi \in L$, $w, f\|\phi\|^{\mathfrak{A}} w, f'$ as usual.

(21) $w, f\|\pi^i\colon\ \phi\|^{\mathfrak{A}} w_{f'(\pi^i)}, f'$ iff $f \subseteq_{\pi^i} f'$, $S_{w_{f'(\pi^i)}}(f'(\pi^i), \|\phi\|^{\mathfrak{A}})$, and
$C_i(w_{f'(\pi^i)}, f') = \{(w'', g)\colon\ \exists (w', f') \in C_i(w, f)$
$(w', f'\|\phi\|^{\mathfrak{A}} w'', g \wedge S_{w''_{g(\pi^i)}}(g(\pi^i), \|\phi\|^{\mathfrak{A}}))\}$.

(22) $w, f\|\varepsilon\colon\ \phi\|^{\mathfrak{A}} w_{f'(\varepsilon)}, g$ iff $f \subseteq_{\varepsilon} f'$ and $w_{f'(\varepsilon)}, f'\|\phi\|^{\mathfrak{A}} w_{f'(\varepsilon)}, g$.

(23) $w, f\|R(\rho_1, \rho_2)\|^{\mathfrak{A}} w, f$ iff $(f(\rho_1), f(\rho_2)) \in \|R\|^{\mathfrak{A}}_w$.

(24) For SDRT formulas $\phi, \psi\colon\ w, f\|\phi \wedge \psi\|^{\mathfrak{A}} w', f'$ iff $w, f\|\phi\|^{\mathfrak{A}} \circ \|\psi\|^{\mathfrak{A}} w', f'$.

(25) For an SDRT formula $\phi\colon\ w, f\|\neg\phi\|^{\mathfrak{A}} w, f$ iff $\neg\exists w' \exists f'.\ w, f\|\phi\|^{\mathfrak{A}} w', f'$.

CDUs containing discourse units are treated like EDUs in terms of commitments; complex units composed only of event units are treated like EEUs. If a player i links a DU with a CDU π, then i commits to the content of π under the same conditions as when π is an EDU. However, the CDU π may contain contributions by other players who may not commit to the CDU as a whole or to re-descriptions of discourse moves they made. Our semantics predicts this, as we saw in our discussion of (18) in Section 5.2.

Let us now return to our original motivating example *Scratches*, which involves two dialogue moves, (1) and (2); a head nod; and some scratches s on a wall. We'll suppose that Peter uses familiar compositional principles to derive the EDU π_1 and its associated logical form, an abbreviated form of which (omitting a proper treatment of the presupposition triggers *our daughter* and *her room*, the genitive construction, and the indexical *our* treated as a constant o) is given in (26) below.

(26) $\pi_1^a\colon\ \exists d, r, x, e.\ (sent(e, x, d) \wedge daughter(d) \wedge of(o, d)$
$\wedge\ to(e, r) \wedge room(r) \wedge of(d, r))$

In terms of clause (21) of our semantics, (26) means that Anne's commitment worlds now all verify the information that the daughter has been sent to her room.

Just after her utterance of (1), Anne performs the nodding gesture ε_0. According to clause (22), the gesture changes the world so that the world's history includes the gesture event, but it does not alter Anne's commitments. Intuitively, however, the nod conventionally signals that there is something in the context that needs to be linked to π_1, and so Anne should commit to linking the denotation of the nod with π_1. But while the nod indicates that there is something discursively relevant in the nonlinguistic context that it indicates, it does not tell us what exactly that is or how it should be related to the discourse context. To capture these two points of uncertainty, we begin with the underspecified logical form in (27). First, we introduce an underspecified EEU label, $\varepsilon_?$, which will ultimately be replaced with a discourse referent and an associated content that specifies the denotation of the nod. Second, we capture the underspecified relation between the denotation of the nod and the incoming discourse by adding an underspecified relation, $R_?$, between $\varepsilon_?$ and π_1. (27) also contains information about the nod ε_0 and its relation to its denotation. As our paper is not about gesture, we forego details here (for more details about how SDRT accounts treat gesture see Lascarides & Stone 2009) and simply assume that the nod will be characterized by some content ψ and that it will be related to its denotation via a relation that we call *Indicates* here.

(27) $\pi_0^a : (\pi_1^a : \phi_1^a \wedge \varepsilon_0 : \exists u(gesture(u) \wedge \psi) \wedge$
$\varepsilon_? : ? \wedge Indicates(\varepsilon_0, \varepsilon_?) \wedge R_?(\pi_1^a, \varepsilon_?))$

Because Anne is responsible for the addition of the relation R to the logical form, it follows that she is committed to the content of π_0, given clause (21). With Lascarides & Stone (2009), we assume that however R is resolved, it will be veridical. This means that in committing to π_0, Anne will commit to the contents associated with the terms of the underspecified relation.

We now complete the logical form for π_0. Following Anne's cue, Peter will next notice the scratches on the wall. Peter takes Anne to have committed to the fact that the scratches on the wall are at least part of the denotation of the nod. The gesture's partially specified denotation adds information about the scratches to the content of the discourse and helps resolve $\varepsilon_?$; in particular, we will add a state of the wall, represented as ε_1,[22] and conclude that $\varepsilon_?$ must include at least ε_1.

(28) $\varepsilon_1 : \exists x \exists y.(on(\varepsilon_1, y, x) \wedge wall(y) \wedge scratches(x))$

22 Here we use subscripts to indicate placement in a linear ordering of EEUs and remain agnostic about its relation to the temporal order of events.

While the state ε_1 certainly plays a role in resolving $\varepsilon_?$, however, it is not the whole story. In particular, identifying $\varepsilon_?$ with ε_1 does not by itself support a resolution of the underspecified relation R. Why should just any old scratches on the wall be relevant to the daughter's being sent to her room? To be relevant, ε_1 must first be conceptualized in an appropriate way. This conceptualization comes not from a closer inspection of ε_1 but from expectations created by the discourse context. In particular, it is natural to expect an explanation to follow Anne's mention of the punishment in π_1 (Asher & Lascarides 2003). Therefore, while an interpreter cannot be sure that $R_?$ will be resolved to Explanation, her expectation of an explanation guides her reasoning process and helps her to properly conceptualize the scratches on the wall. By hypothesizing an Explanation, the interpreter accords a high probability to Anne's committing to a particular conceptualization of the scratches as the outcome of a nonlinguistic event in which Peter and Anne's daughter caused the scratches. This inference leads to the construction in (29) of a CDU containing both the causing event ε_2 and the scratched state of the wall and a Result relation between them expressing their causal dependency.

(29) $\varepsilon_3 : (\varepsilon_1 : \phi_{\varepsilon_1} \wedge \varepsilon_2 : (activity(\varepsilon_2) \wedge agent(\varepsilon_2, d)) \wedge Result(\varepsilon_2, \varepsilon_1))$

The complex unit ε_3 can then serve as the denotation of Anne's nod and an explanation for the punishment described in π_1^a, yielding the final logical form:

(30) $\pi_0 : (\pi_1^a : \phi_{\pi_1^a} \wedge \varepsilon_3 : \phi_{\varepsilon_3} \wedge Indicates(\varepsilon_0, \varepsilon_3) \wedge Explanation(\pi_1^a, \varepsilon_3))$

Once we have the logical form in (30), it is straightforward to process the discourse move in (2) and relate it discursively via Background to ε_3, which is available according to Definition 7 of the Right Frontier of a situated discourse structure.

Julie Hunter
IRIT, Université Paul Sabatier
118 route de Narbonne
31062 Toulouse Cedex 09
France
juliehunter@gmail.com

Nicholas Asher
IRIT, CNRS
118 route de Narbonne
31062 Toulouse Cedex 09
France
asher@irit.fr

Alex Lascarides
School of Informatics, University of Edinburgh
10 Crichton Street
Edinburgh, EH8 9AB
Scotland
alex@inf.ed.ac.uk