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. Extending a coherence-based framework accordingly is not a trivial task. It must cope with (at least) two aspects of how linguistic and non-linguistic moves interact. On the one hand, there can be an asymmetric dependence; the non-linguistic events may form a coherent connected structure in the absence of the linguistic moves, but not vice versa. On the other hand, linguistic moves may contribute to the non-linguistic structure; what is said may bring about the occurrence of a non-linguistic event. 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.

Just after she says this, she nods suggestively over her shoulder and her husband, taking her cue, notices some scratches on the wall behind her. He immediately infers

[*] This work was supported by ERC grant 269427.

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, no extant accounts model the particular *way* in which *e* and e' contribute to the content of *Scratches* and *Table*, respectively. 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 non-linguistic event, a focus of research on modelling multimodal demonstrations (Stojnic et al. 2013, Forbes et al. 2015). The non-linguistic events *e* and e' that contribute to the content of the interaction is over; and it isn't currently visible, only its effects are. While the non-linguistic event is semantically related to the linguistic one, the discursive contributions of the non-linguistic 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. In (5), the discourse unit conveying the content of the second sentence, call it π_2, bears two semantic relations in the rhetorical structure for that discourse: first, π_2 furnishes an explanation and thus Explanation(π_1, π_2) holds, where π_1 is introduced by the first sentence; second, Background(π_2, π_3) holds where π_3 is the unit introduced by the third sentence. The interpretation of *Scratches* should furnish the same discourse structure, but where the non-linguistic event e, or some representation of it, takes the place of the unit π_2. No extant theory of discourse structure can currently derive or even make sense of such a discourse structure containing nonlinguistic entities. For one thing, all such theories have thus far assumed that discourse structures of the sort just sketched apply only to *linguistically given* content or text; here e and e' are not given but inferred. Furthermore, the machinery of discourse structure is designed to deliver something more or less syntactic Forbes et al. (2001), and at least highly structured with recursive construction principles. When theories bother to make explicit what discourse structures are, they take them to be logical forms Asher (1993), Asher & Lascarides (2003) that then are subject to model-theoretic interpretation. Our events e and e' don't have anything like an evident syntactic or recursive structure. Interpreters assign them a description or *conceptualization* that endows them with a recursively structured content, but this description is itself inferred in *Scratches* and *Table* from the discourse structure interpreters are building and into which e and e' are eventually integrated. While one might allow for a logical form encompassing nonlinguistic objects, no discourse theory, or indeed any semantic theory, has investigated the kind of co-dependence of discourse structure and event conceptualization we have just sketched in any formal detail. Nor has any theory investigated how such events enter into discourse structures unless they are referents for, or otherwise specify the interpretation of, context dependent expressions in the linguistic units.

Our goal in this paper is in part to extend work on rhetorical structure and to develop a model for discourses like *Scratches* and *Table*, in which nonlinguistic events somehow contribute the contents of entire discourse units. We will also show that despite the intuitive connection between (5) and *Scratches* and (6) and *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 in discourse structure construction and interpretation but also introduce additional steps in the construction procedure like conceptualization that finally 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 exchanges 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 rules governing the shape of allowable discourse structures; second, it provides a dynamic semantics 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, non-linguistic 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 clearly complementary to many phenomena studied in formal and computational approaches to meaning, but that has not, on the other hand, been systematically modelled within these fields. Modelling this interaction will involve looking at two directions of contextual influence: the effects of non-linguistic 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.

2.1 From nonlinguistic events to the interpretation of discourse

While it is widely recognized that the nonlinguistic context can affect the interpretation of discourse, detailed study in formal semantics of the mechanisms underlying nonlinguistic context dependence has focused largely on how the nonlinguistic context affects the interpretation of certain expressions. This includes work on deictic expressions (Kaplan 1989) and anaphoric reference (Kamp & Reyle 1993) (including deep anaphora: Hankamer & Sag 1976), as well as work on how nonlinguistic topic situations or contexts might influence the interpretation of quantificational expressions, vague expressions, subjective expressions, and so on (Kamp 1981, Stanley & Szabo 2000, Elbourne 2005). Similarly, 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 also yielded computational systems that can automatically find extra linguistic referents for referring descriptions (Kranstedt et al. (2006), *inter alia*). 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).

Examples like *Scratches* and *Table* manifest a kind of contextual interaction that these theories and approaches were not designed to handle. Neither e nor e' serve to supply the referent of a particular lexical item, either through deixis or anaphora; nor do they influence the interpretation of a definite or quantificational expression in the way that a topic situation might. They also are not denoted by nor are they a demonstration of the content of an entire linguistic phrase, and thus they signal a dimension of multimodal meaning that's missing from current HRI systems.

That said, the dependency involved in these discourses is clearly similar to that involved in the purely linguistic variants, (5) and (6). In fact, we believe that the mechanism underlying the context sensitivity of *Scratches* and *Table* is the mechanism of rhetorical dependence already familiar to researchers working on discourse structure. In *Scratches*, we understand e as providing an *explanation* of the punishment described in (1). This relation in turn affects the interpretation of the discourse in ways well-known from work on rhetorical relations: for example, because e provides an explanation of the punishment, it must have taken place before the punishment. This yields a temporal reference point prior to that given by (1), which is then picked up by (2). A nonlinguistic event that plays a different rhetorical role will make a correspondingly different contribution to the temporal interpretation of a discourse: we understand e' as occurring *after* (some part of) the event of moving the table introduced by (3) because we understand it as a *result* of this event.

If we follow this analogy through, however, we come to the conclusion that e not only affects the interpretation of *Scratches*, but its very structure or logical form. That is, e must be contributing the content of an entire *discourse unit*—an instance of a proposition—to the content of *Scratches*. 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. We would end up either relating (2) directly to (1), which would validate the incorrect inference that Anne was cooking dinner when her daughter was being punished, or leaving the two contents unrelated, which would undermine all extant definitions of coherent discourse.

Given these observations, we can formulate a general hypothesis (H) about the way that e affects the interpretation of *Scratches* (and likewise for e' and *Table*):

> (H) A nonlinguistic event e can affect the interpretation of a discourse by contributing the content of an entire discourse unit, 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 sig-

nalling its relevance. Its relevance is then inferred through the kind of reasoning used to infer rhetorical connections between linguistically expressed contents.

In pursuing (H), our work complements that of others who have claimed that rhetorical relations play a key role in the semantic representation of multimodal interaction. Lascarides & Stone (2009) posit that co-verbal 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. They use these connections to model how utterances can affect a salience ranking of nonlinguistic entities. Our study expands upon this work by examining nonlinguistic events that are neither gestural nor co-verbal, and we consider a much wider range of interactions between non-linguistic events and the global discourse interpretation. In addition, 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 full story of the context sensitivity of *Scratches* and *Table* will also require a model of information flow in the opposite direction: from discourse to the interpretation of the nonlinguistic context. Nonlinguistic eventualities that are observed or inferred from the nonlinguistic context cannot serve directly as arguments to rhetorical relations, because these relations take instances of propositions (a proposition paired with a discourse and its context) as arguments (Asher 1993). We will therefore need to model the mechanism by which nonlinguistic events come to be associated with particular semantic contents.

On the basis of perception and world knowledge alone, interpreters will be able to conceptualize many objects and events in their surroundings. Language, however, can guide an interpreter towards a particular conceptualization of these entities, so as to satisfy discourse constraints. In visual illusions like the duck-rabbit illusion, made famous by Wittgenstein, two conceptualizations may compete for our attention (see Figure 1). The linguistic context given by *Do you see the rabbit?* asks an interpreter to conceptualize the figure as a rabbit, ignoring the image of the duck, while the context given by *Do you see the duck?*, asks the interpreter to adopt the competing conceptualization.

Similarly, the way in which a speaker exploits a nonlinguistic event can influence the set of concepts that an interpreter takes to characterize that event. Adding

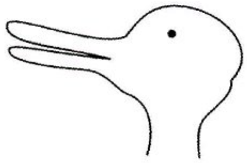


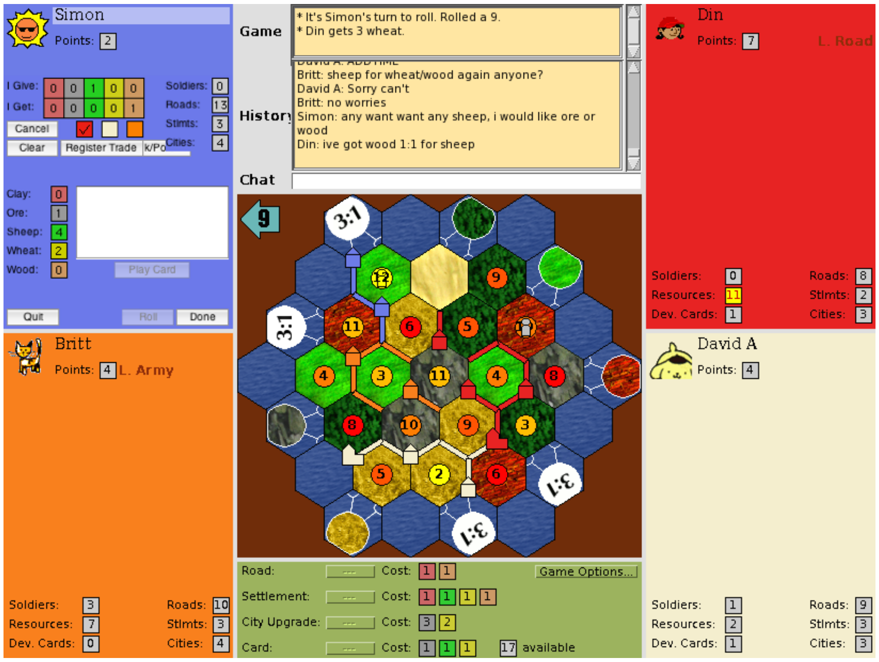
Figure 1: The duck-rabbit illusion

rhetorical links can force an interpreter to *re*-conceptualize a nonlinguistic event in order to properly integrate it into a coherent and meaningful representation of the interaction. 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 reasoning about discourse coherence requires two different interpretations of that context. In fact, because neither e nor e' is 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.

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, although moving to nonlinguistic events greatly complicates the latter task, 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. In this paper, our focus will be primarily on constructing a model for the first task in the setting of situated discourse. We will, however, have some things to say about the second. Our goal is not to ignore it entirely, but to develop a solid model of the first task so that it can be used as a foothold in future research on the second. Our work thus complements that of Larsson (2013) and Dobnik et al. (2013), 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.

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



Simon

Points: 2

I Give: 0 0 1 0 0

I Get: 0 0 0 0 1

Soldiers: 0

Roads: 13

Stlmts: 3

Cities: 4

Clay: 0

Ore: 1

Sheep: 4

Wheat: 2

Wood: 0

Game

* It's Simon's turn to roll. Rolled a 9.
* Din gets 3 wheat.

History

David A: ADD TIME
Britt: sheep for wheat/wood again anyone?
David A: Sorry can't
Britt: no worries
Simon: any want want any sheep, i would like ore or wood
Din: ive got wood 1:1 for sheep

Chat

Din

Points: 7

Soldiers: 0

Resources: 11

Dev. Cards: 1

Roads: 8

Stlmts: 2

Cities: 3

Britt

Points: 4 L. Army

Soldiers: 3

Resources: 7

Dev. Cards: 0

Roads: 10

Stlmts: 3

Cities: 4

David A

Points: 4

Soldiers: 1

Resources: 2

Dev. Cards: 1

Roads: 9

Stlmts: 3

Cities: 3

Road: Cost: 1 1

Settlement: Cost: 1 1 1 1

City Upgrade: Cost: 3 2

Card: Cost: 1 1 1 17 available

Figure 2: A snapshot of the game board for Settlers of Catan

an online version of the game *The Settlers of Catan*. 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 2, 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 the robber to a hex of her choice and steal a resource from a player with a building on that hex. The hex occupied by the robber will generate no resources, regardless of the dice rolls, until it is moved.

Figure 2 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 2 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 2 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 2 (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. (7) is one example:

(7)

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. tomas.kostan 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 (7) are indicated in the left column.[1] Each turn is also identified with an agent, as shown in the middle column of (7). For chat moves, the agent is the player who typed the chat message (e.g., GWFS for turn 434);[2] 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 (7), 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

[1] 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.

[2] Small capitals indicate names that have been abbreviated to save space.

comments on the theft itself or on GWFS' comments in 437 (note that *it* in 439 is an example of a well-established interaction between language and context, since it denotes the wheat, a salient part of the non-linguistic context). Immediately after, in 441, LJAY comments on the result of the theft: that William has 13 resources.

(7) 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 (7) 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 (7) was such an orphan.

These observations triggered a second round of annotations in which annotators had access to both chat moves and game events. A comparison of the first round of annotations with the current annotations shows a significant reduction in orphaned discourse units of over 95% (there were 886 orphans out of a total of 10055 linguistic moves in a sample of 32 games of the corpus). In addition, it shows that many (2006) of the 9879 discourse connections in our sample that were posited based on linguistic-only information have been judged incorrect now that annotators have access to game moves (for more details, see Hunter et al. 2015b). We also found that adding the game events affects judgments about the structure of the chat-only moves. 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 1221 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. (8) illustrates a common pattern in which linguistic moves are grouped together because they jointly cause a nonlinguistic event.

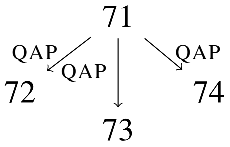
(8)

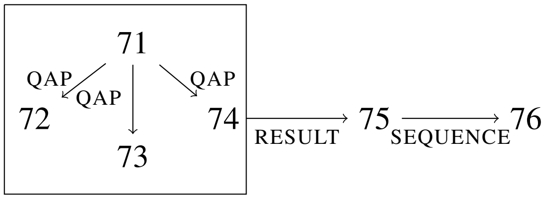
71	tomas.kostan	anyone can offer any wood?
72	william	sry no
73	GWFS	sorry - more 6s and I can oblige then :)
74	LJAY	move the robber and sure :p
75	Server	tomas.kostan traded 3 sheep for 1 wood from a port.
76	Server	tomas.kostan built a road.

Intuitively, neither 72, 73, nor 74 is *independently* responsible for tomas.kostan's decision to trade from a port.[3] 75 is rather the result of his entire failed trade attempt;

[3] A port is a particular location on the Catan board where players can trade under certain circumstances.

that is, the first argument of the inferred Result to 75 is a *complex* discourse unit composed of 71-74.[4] In this way, adding the game events has shed light on the discourse structure of even the chat-only moves in our corpus. 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, that serves as the first argument to the Result relation. QAP stands for the relation Question-Answer-Pair.





(7) 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 game events.[5] This makes our corpus particularly interesting for researchers, including computational linguists and researchers working on human-robot interactions, 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). Moreover, we must be able to assign individual events an appropriate typing or interpretation. As mentioned in Section 2, the event *e* in *Scratches* has to be thought of in a particular way: Peter must understand it as an event in which his daughter scratched the wall (or was in some other way responsible for the scratches). It is not enough for him to realize that *something* happened to the wall or that *someone* scratched it. Nor is it sufficient for him to simply recognize that there is a mark on the wall. Deriving *e* and its appropriate interpretation is an inferential process, but one that is not yet well enough understood to be employed in the segmentation and annotation of, say, a video-based corpus.

Our Settlers corpus largely circumvents the problem of how nonlinguistic events are individuated and appropriately typed or conceptualized. This is partly due to its

[4] Asher et al. (2016) describes the SDRT-inspired annotation of the corpus. For more on SDRT representations see section 5.1.

[5] 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 (9) in Section 4.1.

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 they are also able 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 harder to control, such as somebody eating popcorn too loudly and eliciting a loud *Shhh!*. 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 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 doesn't 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 built 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; while the messages are helpful as a record for annotators and for players who might have a lapse in attention, 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 by using information about the dice rolls and where the Red player has built settlements; and so on. Whatever argument one might make that the game events are linguistic because they are linguistically described in the Game window could be directly adapted to argue that the events are nonlinguistic because they are visually represented by the game interface.

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 (7) is that while we understand (7) 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 contents of coherence relations to the content of (7), this is not their *raison d'être*. The agents clearly do not make their game moves with the intention of contributing to a discourse, but 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 the nonlinguistic events involved in these examples and reexamining standard assumptions about coherence relations. In section 4.1 we introduce a distinction between discourse internal and discourse external events which is relevant for situated discourse, but which cuts across the linguistic/nonlinguistic distinction, and we compare discourse internal/external interactions with a certain kind of multimodal interaction. In section 4.2, we apply the discussion of discourse external events and multimodal interactions to an old discussion about the role of intentions in interpretation.

4.1 Discourse external events and structural asymmetries

In a situated discourse d, there might be many things going on in the larger situation whose existence is independent of d. 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 what we will call *discourse external* events. These are events that take place in the context of d, or whose result states are apparent in that context, but that are not constrained to be relevant to d.

Discourse external events contrast with *discourse internal* events. A discourse internal event is an event whose occurrence depends on the discourse in the sense that we understand that its *raison d'être* is to contribute semantic content to d. It is tempting to define discourse internal events in terms of intentions: where c is the semantic content associated with an event e (i.e., c describes e), e is internal to a discourse d just in case its agent performed e with the intention of contributing c to d. This definition, however, excludes meaningful gestures and facial expressions that interlocutors make unconsciously. Such events are not intentional, though arguably they are understood as existing in order to contribute to the speaker's larger message. We thus prefer a counterfactual classification of a discourse internal event *e* as an event that would not have occurred if its agent had not been contributing to *d* and which is such that the content of *d* would have changed had *e* not occurred.

Our hypothesis (H) allows a discourse external event to contribute to the content of a discourse *d* in the same way that a discourse internal event can. It is of course well known that a discourse external event can affect the interpretation of *d*. Suppose you are outside and you look up in the sky and see some people parachuting down from high above you. You say, "I never want to do that!". In this case, the interpretation of *that* depends on the discourse external event that you and your interlocutors (if there are any) are witnessing. But (H) pushes the potential contribution of a discourse external event *e* to the content of of the discourse *d* much further: it says that *e* can contribute an entire proposition to *d* and that in so doing, it can change the very structure or logical form of *d*. In this case, its role is not to satisfy the semantic demands of certain expressions, and its influence cannot be modelled as a mere step in evaluating the semantic representation of *d* against the world or a model.

When a discourse external event *e* is woven in to a discourse, it will in general also simultaneously constitute a part of another dynamically evolving structure to which it is its *raison d'être* to contribute. This leads to asymmetric discourse configurations, which we will call *asymmetric dependencies* in discourse structure. A discourse structure with an asymmetric dependency contains a substructure that we call a *core* that develops independently of the larger structure, but the development of the larger structure depends on the development of the substructure. Let's return to (7). William didn't roll the dice in turn 435 because GWFS said *noooo!* in 434; in fact, there is no constraint that the dice roll be relevant to anything that has been said in the chat. William's move is external to the chat as its *raison d'être* is to contribute to the game—the players are in a game state where for the game to progress forward, GWFS must roll the dice. The players can nevertheless exploit or *appropriate* game moves, as they do with 433.0.4, in order to build a larger structure that includes both game events and chat moves. This gives rise to an asymmetric dependence because while the game moves form a coherent and connected structure of their own, the chat moves, were we to remove the contributions from the game state, would not.

When a discourse structure is appropriated and woven into a larger structure, this process can lead to a re-conceptualization of game events analogous to the regrouping of chat events that we observed in example (8). Nevertheless, the substructure retains a kind of independence because the resulting conceptualization must be at least supported by the information in the substructure alone. Consider (9):

(9)	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

GWFS's utterance of *touché* requires a certain parallel structure between LJAY's move and his prior move. His utterance is justified not by a single event, but by a certain relation between the set of moves 154.1-154.3 and the set 159.1-159.4. Thus interpreting (9) requires us to add information beyond what is strictly required by the game structure alone. Still, the substructure of game events yields a complete discourse structure that is consistent with this reconceptualization.

The distinction between discourse internal and external events cuts across the linguistic/nonlinguistic distinction: that is, it is possible to have a substructure in an asymmetric dependency that consists entirely of linguistically-specified discourse moves. This can happen when conversational participants overhear and comment on, for example, another conversation or a public speech.

In multimodal discourse in which a nonlinguistic event can contribute a whole discourse unit, we find, in addition to asymmetric structures, what we will call *interleaved structures*. One frequent case in our corpus is when a trade negotiation leads to a nonlinguistic trade, as in (8). In this structure, as opposed to the asymmetrically dependent structure in (7), the chat moves bring about a change in the game state. The connections between the chat moves and game events are very much like the connections that we are used to seeing in purely linguistic discourse.

(10) provides another, more subtle example that shows that interleaved structures are nevertheless different from structures for purely linguistic discourse, at least as they 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 to trade with LJAY, but his efforts have proven futile. Despite the warning in 538 that it might be a good idea for her 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 what he said. 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 plan B. 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. As such, the explanation that we infer, which is a meta-level relation concerning GWFS's strategy, is external to the game structure. On the other hand, the warning is a game move just as any proposed but ultimately unsuccessful offer is; so the warning itself is *internal* to the game structure. Thus even internal moves in an interleaved structure can give rise to asymmetric structures, when these moves play a kind of double role.

4.2 The hermeneutical stance

The data we have considered together with hypothesis (H), introduced in Section 2.1, have led us to countenance the existence of asymmetric dependencies and interleaved structures. In this section, we consider how these structures lead us to question 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 m' 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 external events and 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 external 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 allowing 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.

This has general consequences for the way we think of discourse. In our view, semantic structures composed entirely of what are traditionally classified as discourse moves (including, perhaps, discourse internal 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. 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.

(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

The claim that the resource distribution should be construed as a result of the dice roll pushes the argument against communicative intentions even further. Not only are *communicative* intentions not responsible for the resource distribution, *no* intentions are, at least not directly. The distribution stems from the rules of the Settlers game, but it is semantically relevant in a way that requires us to interpret it as contributing to a larger, asymmetric structure.

5 Technical changes to discourse structures

What differentiates the representation of a purely linguistic discourse from an asymmetric or interleaved structure is not the nature of the connections in the representation nor the nature of the entities connected. What differs is the shapes of these structures, their model theory, and the types of inferences that underly constructing these structures during discourse parsing. In this section, we take a look at how to modify structural constraints that have been presented for linguistically-expressed discourse and then discuss changes to the model theory needed to interpret situated discourse structures. 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—a discourse move is coherent in the context of a discourse only insofar as it is coherently related to some other part of that discourse. Details of the formal implementation in SDRT can be found in the Appendix.

The nodes of our weakly connected graphs will be either (i) *elementary discourse units* (EDUs), which we will take to be the contents of linguistically-specified clauses, (ii) *elementary event units* (EEUs), which are contents assigned to nonlinguistic events in the context, or (iii) *complex discourse units* (CDUs), which can be composed of discourse units (DUs), event units (EUs), or a combination of the two. The directed edges or arrows in our graphs are labelled with names for discourse relations. Following SDRT and Polanyi (1985)[6] we distinguish *subordinating* edges, which are labelled with names for subordinating relations such as Elaboration, Explanation,

[6] Rhetorical Structure Theory (Mann & Thompson 1987) also has a similar distinction.

and Background, from *coordinating* edges, which are labelled with coordinating relations such as Contrast, Narration (Sequence)[7], and Result.

5.1 Structural constraints on situated 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. We turn now to a discussion of the consequences of asymmetric and interleaved structures for discourse structure.

5.1.1 Asymmetric structures

In a purely asymmetric structure, the substructure has an autonomous existence; units from the larger structure that are not in the substructure can attach via rhetorical relations to units in the substructure, but they do not change the substructure (except by possibly adding CDUs, as explained in Section 4.1). To make asymmetric structures more precise, we first define a discourse graph in SDRT.

Definition 1 *A discourse G is a tuple $(V, E_1, E_2, \ell, Last)$, where V is a of* EDU*s and* CDU*s, E_1 a set of edges representing discourse relations E_2 a set of edges relating* CDU*s to their members, ℓ a function from elements of E_1 to discourse relation types, and Last a label for the last unit in V relative to textual order.*

Let $e(x,y)$ mean that the edge e connects its initial point x to its end point y. We also say that a finite set X of units in a graph G form a *rope* just in case the transitive closure of the edges in G over X induce a transitive, asymmetric ordering R over X in which every element a not identical to an endpoint o is such that $R(a,o)$ or $R(o,a)$. Note that any chain over X, defined in the standard way, is a rope, and every set of chains over X all with the same endpoints forms a rope as well.

Definition 2 *An asymmetric structure $\mathscr{S}$ is a graph with a maximal rope $\mathscr{C} \neq \mathscr{S}$ with $\mathrm{Last}^{\mathscr{S}} \in V^{\mathscr{C}}$ and such that there are no edges $e \in E_1^{\mathscr{S}}$ such that $e(a,b)$, $a \in V^{\mathscr{S}} \setminus V^{\mathscr{C}}$ and $b \in V^{\mathscr{C}} \cup \{\mathscr{C}\}$.*

The substructure $\mathscr{C}$ in the asymmetric structure $\mathscr{S}$ defined in definition 2 is the *core* of $\mathscr{S}$. Such a core is typically represented as a CDU in SDRT. Typically, we are interested in thematically unified ropes such as one that contains the events

[7] 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.

that directly affect the flow of the game in our corpus or one that contains the contributions of a set of speakers and where the rest of the conversation is a series of comments on that rope. According to our definition, however, a core with a particular theme will contain any non-thematically related material that ends the conversation. This means that the core may contain in effect several themes; alternatively, we might judge material at the end of the conversation to be outside the core on thematic grounds.

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 them *danglers* Venant et al. (2013). Units contributed by appositive relative clauses, for example, generally function as danglers. A core in an asymmetric structure $\mathscr{S}$ is like the backbone of $\mathscr{S}$, and nodes appearing only outside that core are danglers. But they are danglers with a twist: there can be a coherence relation between two danglers that extend from separate nodes of the core, which leads to non-treelike, "rectangular" discourse structures not normally countenanced in coherence-based theories of discourse (indeed, RST Mann & Thompson (1987) is restricted to trees). Consider (12), which is a fragment of a negotiation dialogue (with a last turn of a nonlinguistic nature (X ended his/her turn):

	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.
(12)	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.

(12) yields an asymmetric structure depicted in Figure 3 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 discourse 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 on the dice roll in 341 or the CDU containing the dice roll and the distribution (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 (Polanyi 1985, Hobbs 1985, Kehler 2002, Asher 1993), it must

be related to 344 so as resolve the linguistically implicit arguments to the relation *better* to their intuitive values. This yields the following rectangular structure:

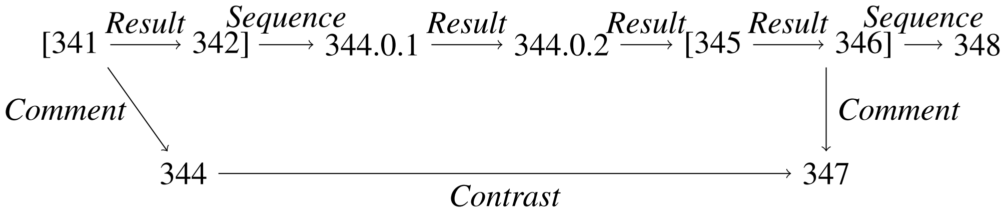


Figure 3: The discourse graph for (12)

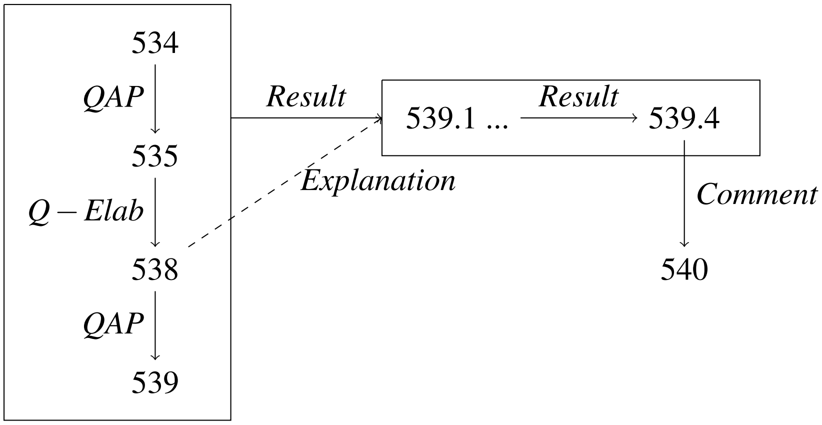
The following example illustrates the same point, but with a different relation between the danglers and an explicit discourse connective. Note that we have not included the full game sequence of events that form the core of the structure for this dialogue.

(13)	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

(13) is a semi-constructed example. In the original, GWFS says, “also you keep thieving me" and he does not make the comment in 238. Still we find (13) 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.

While asymmetric structures frequently involve danglers, a different kind of asymmetric configuration occurs in (10). The failed trade negotiation (turns 534-539) results in GWFS’s playing a Soldier card and stealing from LJAY (539.1-539.1). As discussed earlier, the chat moves and the game moves are a part of the game. There is a Result from the linguistic negotiation to the Soldier card sequence; so 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 external to the game structure. The graph below gives the representation of (10); the dashed arrow represents the game-external arrow.



This is another example of an asymmetric configuration that is unfamiliar from work on rhetorical structure.

5.1.2 Discourse salience and the right frontier

Rectangular structures like that given by (12) 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.[8]

Normally, a coordinating relation such as Sequence or Result is understood as pushing the discourse forward, thereby shutting of 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 the node *Last*—which is the node most recently attached to the discourse graph—as well as any unit that is superordinate to a unit on the RF through a subordinating relation and any CDU that includes a node on the RF.[9] Definition 3 formalizes this definition, using the definition of a discourse graph in Definition 1. Definition 3 defines those

[8] 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*.

[9] 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.

nodes x on the RF of an SDRT graph G, written $\textsc{rf}_G(x)$, which are accessible for the next unit to rhetorically attach to:

Definition 3 *Let $G = (V, E_1, E_2, \ell, Last)$ be a discourse graph. $\forall x \in V$, $\textsc{rf}_G(x)$ iff* ***(i)*** *$x = Last$; or* ***(ii)*** *$\exists y \in V$ $\textsc{rf}_G(y)$ and $\exists e \in E_1$, $e(x,y)$ and Sub(e); or* ***(iii)*** *$\textsc{rf}_G(y)$ and $\exists e \in E_2$, $e(x,y)$*

The definition of the RF entails that attaching to the RF via a subordinating 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. This doesn't happen in (12); despite the Result and Sequence relations intervening between 342 and 345, the comment in 347 is still interpretable as standing in a Contrast relation with 344, producing the rectangular structure observed above. There is a certain level of independence between the core $\mathscr{C}$ of an asymmetric structure and the larger structure in which it is embedded.

Examples like (9), which are frequent in the *Settlers* corpus, illustrate another type of violation of Definition 3: a nonlinguistic event occurring after another nonlinguistic event does not always shut off access to the earlier move. In (14), turns 155 and 157, which are attached via Sequence, should render 154.1-154.3 inaccessible, yet 158 is understood as a Comment on the earlier moves. Likewise, 163 is understood as a Comment on 159.1-159.4 despite the intervening moves.

(14)

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, constraints on chat moves are not completely independent of the structure of the game moves. Consider the minimal variant (15) of (14).

(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 much more difficult to return to the prior Soldier card event, thereby making 163 less coherent.

The discussion of (14) 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. An interaction like the following constructed example would seem to be possible (Hunter et al. 2015a: see also Ginzburg (2012)).

(16)	100	tomas.kostan	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.
	102'	tomas.kostan	What if I offer you two ore?

Had GWFS made only the move described in boldface in 101, his move in 102 would have been coherent. Once he makes the move described 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' is perfectly felicitous even though it builds directly off of the boldface move, ignoring the italicized move. This is, we think, because tomas.kostan is not immediately committed to the discourse structure that GWFS builds: he can effectively ignore some aspects of GWFS' prior commitments when making 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 he 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, we combine this general constraint on discourse with our observation of rectangular structures described in Section 5.1.1. Adding EEUs and CDUs containing EEUs to the set V, we propose the following hypothesis: any DUs, EUs, and CDUs in the core $\mathscr{C}$ of an asymmetric structure will remain accessible to new moves in the larger structure so long as no commentary is made on the moves from $\mathscr{C}$. In addition, all of the nodes on the RF of the preceding discussion will also remain accessible. Once the content of a linguistic move m is attached to a node n from $\mathscr{C}$, however, then the RF of the preceding discussion disappears, unless m attaches to it as well, and all nodes in $\mathscr{C}$ that should be inaccessible from n according to Definition 3 also become inaccessible to m.

Definition 4 below formalizes these observations. Let $Acc(G)$ be the set of labels on the RF of a graph G as in Definition 3, $End(e)$ be the endpoints of an edge e, and $G - G' = (End(E_1 \setminus E_1'),\ E_1 \setminus E_1',\ E_2 \setminus E_2',\ \ell \restriction (E_1 \setminus E_1'),\ x)$, with x the last element in $V \setminus V'$ ordered by a linear ordering $\prec \subseteq V \times V'$. [10]

Definition 4 *let* $G^s = (V^s, E_1^s, E_2^s, \ell^s, Last^s)$ *be an asymmetric discourse graph and* $G^{c_s} = (V^{c_s}, E_1^{c_s}, E_2^{c_s}, \ell^{c_s}, Last^{c_s})$ *be the core of* G^s. *Then:* $\mathrm{RF}(G^s) \ = Acc(G^s - G^{c_s}) \cup \{u \in V^{c_s} : \neg\exists e \in (E_1^s \setminus E_1^{c_s})\ \exists y \in V^{c_s}(y \in End(e) \wedge u \prec y)\}$

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. 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.

Definition 4 effectively means that people make EEUs salient, or not, 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.

5.1.3 Interleaved structures

While interleaved structures will be subject to the RFC defined in Definition 4, sometimes, for a move to coherently contribute to an interleaved structure, certain semantic constraints will need to hold as well. (17) illustrates these limitations.

[10] The Settlers corpus satisfies the presumption in Definition 4 that the units are in a total linear order, but multimodal conversation is generally non-linear. We eschew this complexity here.

(17)

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

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 to make a suboptimal play, 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 (e.g., dmm cannot now accept ljay's offer because she's ended her turn).

This observation brings out a semantic point deeper than (H). (H) 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 the classic work on bargaining Osborne & Rubinstein (1990)). (17) shows that even this kind of modular approach to interaction is too simple. 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. We turn now to a discussion of these mechanisms.

5.2 The semantics of situated discourse structures

We have argued that EEUs can stand in the same coherence relations that EDUs do, and that there is often effectively no difference between what an EDU contributes rhetorically to a dialogue and what an EEU contributes. We capture this equivalence at the level of logical form by featuring the same coherence connections between DUs and EUs and by associating each EDU and each EEU with a logical representation of its content. The depth of interaction between nonlinguistic and linguistic events shows that the standard tools for discourse interpretation, however, will need to be rethought.

In classic dynamic semantics, a context or information state is a set of worlds, or of variable assignment assignment functions, or a set of world-assignment pairs, or a set of such sets (Groenendijk & Stokhof 1991). 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 C induce a monotone decreasing function on the world components C_w of those contexts $f : C_w \to C_w$ such 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 dynamic modal operators (Venant & Asher 2016), to give just two examples.

While these accounts provide a rich and powerful arena for defining how a discourse update with a dialogue move affects 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. To put this another way, 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. Speakers can even be inconsistent, though in that case our function f reaches a fixed point and no information growth is possible.

While we agree that speakers are free to say whatever they want, 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 (17), he cannot give any clay to LJAY later on, unless he first gets more clay. If dmm wins the game, then no one else can win the game.[11]

This affects possible continuations off of the contents of speech acts. Before dmm trades his clay away, he could have responded to LJAY's answer with 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, this possible continuation of her linguistic move is no longer possible. All they can do now is talk about her offer; they are no longer free to act on it. The world isn't just there for us 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 figure in representations of multimodal interactions, as we have proposed they should 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.[12] The actual world allows for certain possible futures, in virtue of which we decide to act. But once we act, some of those possibilities are now 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 $\mathscr{L}$ that constrains how histories can develop; in a model of *Settlers* for instance, $\mathscr{L}$ 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 that links each DU or EU to a content.[13]

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 5.1.2).

[11] 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.

[12] 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, but we will eschew details here.

[13] Note that DUs express propositions that are not about the speech acts these DUs themselves denote, and so we cannot make use of Davidson (1968/69)'s proposal about action sentences to handle this linking.

A model $\mathfrak{A}$ for a graph G representing a conversation with n players will then be a tuple $\mathfrak{A} = \langle D, W, C_1, \ldots, C_n, S, \mathscr{L}, \mathfrak{h} \rangle$, where D is the domain of individuals and events including a set of agents, W is the set of worlds, and for each agent i, C_i is an accessibility relation. Finally, we set the function $\mathfrak{h} \colon W \to (\mathscr{P}(D))^*$, 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.

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 '$\varepsilon : \phi$' mean that ε has content ϕ. Then an EEU $\varepsilon : \phi$, 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 $\mathscr{L}$ and $f'(\varepsilon_n) \in \mathfrak{h}_n(w')$ ($f'(\varepsilon_n)$ is included in the set of eventualities at n) and $\mathfrak{h}_m(w) = \mathfrak{h}_m(w'), \forall m < n$, 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 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 $\pi : \phi$ (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 $\pi^i : \phi$. We take a minimal view to commitment change made by discourse actions here.[14] 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 is interpreting a discourse move π by a speaker i 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 that i commits to $R(\rho, \pi)$. 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 $\varepsilon : \phi$, unless: (i) she is causally responsible for ε, or (ii) she makes ε part of the

[14] For a full treatment of nested, higher-order commitments, see Venant & Asher (2016).

discourse structure by relating it to another DU ρ via a discourse relation R. In the 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 the units it connects, or not. Most moves that involve relations to EEUs, like Result, Explanation and QAP, are *veridical*, which means they entail the dynamic conjunction of the contents of their arguments. Details are in the Appendix.

To illustrate our semantics, let's return to (17). dmm's turn 123 introduces a question that commits him to 2 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 can't trade with LJAY even though he intended to trade with someone and LJAY was willing.

The discourse structure partially models what happens, if turns 123-126 form a CDU that then results in 127.0.1. The semantics says that the two negative responses to the trade offer initiate the nonlinguistic action. It also implies that dmm commits to the negative responses by GWFS and inca as well as to the result relation with the bank trade offer. But this also means that dmm *doesn't commit* to LJAY's response in 127. We note that given the way the semantics is set up in the Appendix, neither 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 evolution and interpretation. It complements work on multi-party dialogue that has compared features of multi-party dialogue with monologue

(Ginzburg & Fernández 2005) and explored the behavior 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.[15] These participants may eavesdrop on d and then exploit elements of d in their own conversation; for their conversation, moves in d are discourse external, 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 work 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 and Stone argue, for instance, that one cannot coherently use Contrast to connect a gesture to its co-verbal speech, while our corpus has many instances of EEUs like nonlinguistic offers connecting with Contrast to a retort like *But I don't have any sheep* as in (18):

(18) 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 non-linguistic moves: i.e., 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 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

[15] For further discussion see Dynel (2010).

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 which structures are possible and the model theory for interpreting them.

An alternative to using SDRT to spell out the details of our analysis would be to use a Question Under Discussion (QUD) model, such as Ginzburg (2012). But some motivations for adopting a QUD model don't really apply to our data. Specifically, exploiting question-answer congruence to constrain focus (Halliday 1967, Roberts 2012) isn't relevant when the contents associated with non-linguistic events don't have a focus structure in any obvious way. Further, accounts that take the answering of questions or the seeking of information as the driving force behind dialogue production or interpretation would seem out of place, given our data. That said, QUD, like SDRT, has served to analyze a variety of anaphoric and elided expressions featured in conversation, including an analysis of 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 a kind of anaphoric dependency, which gets resolved by linking them 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 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 contrasts with a common "bottom up" view of how discourse structures are built in other theories, as in 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 we expect information flow 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 (2) shows, 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 embodied 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 conceptualisation of nonlinguistic events—which has facilitated the study of the codependent task of determining how conceptualizations of nonlinguistic events influence the construction of discourse representations. 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 co-verbal 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, Sergos, 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)*, Paris.

Afantenos, Stergos & Nicholas Asher. 2010. Testing SDRT's right frontier. In *The 23rd international conference on computational linguistics (coling)*, 1–9.

Alahverdzhieva, Katya & Alex Lascarides. 2010. Analysing language and co-verbal gesture in constraint-based grammars. In S. Müller (ed.), *The 17th international conference on head-driven phase structure grammar (hpsg)*, 5–25. Paris.

Asher, Nicholas. 1993. *Reference to abstract objects in discourse*. Kluwer Academic Publishers.

Asher, Nicholas, Julie Hunter, Mathieu Morey, Benamara Farah & Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *The tenth international conference on language resources and evaluation (lrec 2016)*, Paris, France: European Language Resources Association (ELRA).

Asher, Nicholas & Alex Lascarides. 2003. *Logics of conversation*. 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.

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*, Washington, DC.

Clarke, Alasdair, Micha Eisner & Hannah Rohde. 2015. Giving good directions: Order of mention reflects visual salience. *Frontiers in Psychology* 6(1793).

Davidson, Donald. 1968/69. The logical form of action sentences. In Peter Ludlow (ed.), *Readings in the philosophy of language*, 337–346. MIT 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. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-41578-4_5. 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://mediazioni.sitlec.unibo.it, ISSN 1974-4382.

Elbourne, Paul D. 2005. *Situations and individuals*, vol. 90. Mit Press Cambridge.

Elsner, Micha & Eugene Charniak. 2011. Disentangling chat with local coherence models. In *Proceedings of the association for computational linguistics (acl)*, Portland, Oregon.

Forbes, Katherine, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Aravind Joshi & Bonnie Webber. 2001. D-LTAG System – discourse parsing with a lexicalized tree adjoining grammar. In *The esslli-01 workshop on information structure, discourse structure and discourse semantics*, .

Forbes, Maxwell, Rajesh Rao, Luke Zettlemoyer & Maia Cakmak. 2015. Robot programming by demonstration with situated spatial language understanding. In *Proceedings of icra*, .

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, New Hampshire, USA.

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

Ginzburg, Jonathan & Raquel Fernández. 2005. Scaling up from dialogue to multilogue: some principles and benchmarks. In *The 43rd annual meeting on association for computational linguistics*, 231–238. Association for Computational Linguistics.

Ginzburg, Jonathan & Ivan A. Sag. 2001. *Interrogative investigations: The form, meaning and use of english interrogatives*. CSLI Publications.

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

Groenendijk, J. & M. Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy* 14. 39–100.

Groenendijk, Jeroen. 2003. Questions and answers: Semantics and logic. In *The 2nd colognet-elset symposium. questions and answers: Theoretical and applied perspectives*, 16–23.

Halliday, Michael AK. 1967. Notes on transitivity and theme in english: Part 2. *Journal of Linguistics* 3(02). 199–244.

Hankamer, Jorge & Ivan Sag. 1976. Deep and surface anaphora. *Linguistic inquiry* 7(3). 391–428.

Hobbs, Jerry R. 1985. On the coherence and structure of discourse. Tech. Rep. CSLI-85-37 Center for the Study of Language and Information, Stanford University.

Hobbs, Jerry R., Martin Stickel, Douglas Appelt & Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence* 63(1–2). 69–142.

Hunter, Julie, Nicholas Asher, Eric Kow, Jérémy Perret & Stergos Afantenos. 2015a. Defining the right frontier in multi-party dialogue. *SEMDIAL 2015 goDIAL* 95–103.

Hunter, Julie, Nicholas Asher & Alex Lascarides. 2015b. Integrating non-linguistic events into discourse structure. In *The 11th international conference on computational semantics (iwcs)*, 184–194. London.

Joty, Shafiq, Giuseppe Carenini & Raymond Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics* .

Kamp, Hans. 1981. The paradox of the heap. In *Aspects of philosophical logic*, 225–277. Springer.

Kamp, Hans & Uwe Reyle. 1993. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Kluwer Academic Publishers.

Kaplan, David. 1989. Demonstratives. In J. Almog, J. Perry & H. Wettstein (eds.), *Themes from kaplan*, Oxford.

Kehler, Andrew. 2002. *Coherence, reference and the theory of grammar*. CSLI Publications, Cambridge University Press.

Kendon, A. 1983. Gesture and speech: How they interact. In J. Wiemann & R. Harrison (eds.), *Nonverbal interaction*, 13–46. Sage Publications.

Kranstedt, Alfred, Andy Lüking, Thies Pfeiffer, Hannes Rieser & Ipke Wachsmith. 2006. Deixis: How to determine demonstrated objects using a pointing cone. In *Gestures in human-computer interaction and simulation*, 300–311. Springer.

Larsson, Staffan. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation* 25(2). 335–369.

Lascarides, Alex & Nicholas Asher. 2009. Agreement, disputes and commitment in dialogue. *Journal of Semantics* 26(2). 109–158.

Lascarides, Alex & Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics* 26(4). 393–449.

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

Liang, P.2̇005. *Semi-supervised learning for natural language*: Department of Electrical Engineering and Computer Science, MIT dissertation.

Lücking, Andy. 2016. Modeling co-verbal gesture perception in type theory with records. In *Proceedings of the 2016 federated conference on computer science and information systems*, 383–392. IEEE. http://dx.doi.org/10.15439/2016F83.

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.

Moore, Johanna D. & Cécile L. Paris. 1993. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics* 19(4). 651–695.

Muller, Philippe, Stergos Afantenos, Pascal Denis & Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of coling 2012*, 1883–1900. Mumbai, India: The COLING 2012 Organizing Committee. http://www.aclweb.org/anthology/C12-1115.

Osborne, Martin & Ariel Rubinstein. 1990. *Bargaining and markets*. Academic Press.

Perzanowski, Dennis, Alan Schultz, William Adams, Elaine Marsh & Magda Bugajska. 2001. Building a multimodal human-robot interface. *Intelligent Systems* 16(1). 16–21.

Polanyi, Livia. 1985. A theory of discourse structure and discourse coherence. In P. D. Kroeber W. H. Eilfort & K. L. Peterson (eds.), *Papers from the general session at the 21st regional meeting of the chicago linguistics society*, 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 Szabo. 2000. On quantifier domain restriction. *Mind and Language* 15. 219–261.

Stojnic, Una, Matthew Stone & Ernie Lepore. 2013. Deixis (even without pointing). *Philosophical Perspectives* 27(1). 502–525.

Venant, Antoine & Nicholas Asher. 2016. Ok or not ok? commitments, acknowledgments and corrections. *Semantics and Linguistic Theory (SALT 25)* 595–614.

Venant, Antoine, Nicholas Asher, Philippe Muller & Pascal Denis Stergos D. Afantenos. 2013. Expressivity and comparison of models of discourse structure. In *Proceedings of Sigdial 2013*, 2–11. Metz, France.

Webber, Bonnie. 1988. Tense as discourse anaphor. *Computational Linguistics* 14(2). 61–73.

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. 601–713.

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.

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_i 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 EDU, the person responsible is the speaker/author; for a CDU, the person responsible is the person responsible for its last EDU; and for an EEU, the person responsible is the one who committed to that action (where EEUs that lack a volitional agent have no superscript). Situated SDRT formulas are defined recursively in terms of L:

Definition 5 SDRT Formulas:

i. Where ϕ is a formula of L, $\pi : \phi$ and $\varepsilon : \phi$ 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 $\pi : \phi$ is an SDRT formula.

In the semantics of situated SDRT formulas, assignments assign EEU variables denotations from a set of nonlinguistic entities, while EDU variables are assigned speech acts. We treat all labels in SDRT as discourse referents or existentially bound variables. A modal accessibility relation over world assignment pairs models speaker commitments: an agent i's commitment slate 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. For questions, we 'lift' the basic semantics to sets of world-assignment pairs: a question partitions the input set of worlds so that each equivalence class in the output partition corresponds to a possible answer (Groenendijk 2003), though we forego lifting here.

We formalize CCPs for DUs and EUs in (20) and (21), respectively. $f \subseteq_\rho f'$ indicates f' extends f over an EDU or EEU ρ. ρ, ρ_1, ρ_2 range over EEUs and EDUs. To denote the shift in worlds due to an action x being performed at time n, we will write w_x to denote a world such that $\mathfrak{h}_m(w_x) = \mathfrak{h}_m(w)$ for all $m < n$, and $f'(x) \in \mathfrak{h}_n(w_x)$. Update is defined with relational composition $\circ$. Let i be any speaker. Then (19)–(24)

provide an interpretation of DUs (the same interpretation clauses apply to both EDUs and CDUs) and also EEUs and their associated formulas:

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

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

(21) $w,f\|\varepsilon_n : \phi\|^{\mathfrak{A}}w_{f'(\varepsilon_n)},g$ iff $f \subseteq_{\varepsilon_n} f'$ and $w_{f'(\varepsilon_n)},f'\|\phi\|w_{f'(\varepsilon_n)},g$.

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

(23) For SDRT formulas ϕ,ψ: $w,f\|\phi \wedge \psi\|^{\mathfrak{A}}w',f'$ iff $w,f\|\phi\| \circ \|\psi\|^{\mathfrak{A}}w',f'$.

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

Note that the shifting of the world parameter in the evaluation happens not only at the world of evaluation but also in the commitment slate of the speaker. Note also CDUs are treated like EDUs in terms of their commitments. 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 with our discussion of (17) above.

word count (with numbers): 16,388