Message Exchange Games in Strategic Contexts

Nicholas Asher[1]*, and Soumya Paul[2], and Antoine Venant[2]
[1]CNRS, IRIT and [2]Université de Toulouse 3, IRIT

December 28, 2016

1 Introduction

Conversations often involve an element of planning and calculation of how best one can achieve one's interests. We are interested in how conversations proceed in a setting in which dialogue agents cannot assume that their interests coincide with those of their interlocutors, and we think this is a promising starting point for a general model of conversation. While there is a large literature in linguistics and in AI on cooperative conversation stemming from [20], there is little theoretical and formal analysis of conversation in non-cooperative situations. The work of [42], where cooperativity is determined only by the social conventions guiding conversation, obligations that do not presuppose speakers to adopt each other's goals, constitutes an important exception. Still, the formal structure of such conversations remains largely unexplored. We propose here a formal theory of message exchange in settings where agents do not necessarily share interests and goals.

In particular, a little explored element in linguistics is the general "shape" of a conversation, its overall structure and the effects of this structure on content. The goals of conversational participants and the context of moves they have already may explain why they make the subsequent discourse moves they do and give a coherence to the conversation as a whole. For conversations where agents share conversational goals and interests, a broadly Gricean answer explored by [21, 22] *inter alia* is that the discourse is organized around a problem that it is in the common interest of the participants

*Thanks to Julie Hunter, Alex Lascarides, David Beaver, Eric McCready, Daisuke Bekki, Chris Barker, Erich Grädel, Hans Kamp, Benedikt Löwe, Julian Schlöder, Itai Sher, to the participants of the Rutgers Workshop on Coordination and Content and to an anonymous reviewer for the *Journal of Philosophical Logic* for their helpful comments on previous versions of this paper. This work was supported by ERC grant 269427.

to resolve; the structure of the conversation reflects the structure of the decision problem, or rather the reasoning of conversational participants to construct a plan that solves the common decision problem.

Strategic conversations, conversations where cooperativity or shared goals and interests cannot be assumed, do not instantiate reasoning about a common decision problem. But they do have a structure that is shaped by the goals and interests of the participants. Strategic conversations are also reactive: to achieve their goals, each participant needs to plan for anticipated responses from other participants. Assuming that conversationalists are rational, what they say and how they interpret what is said should follow as actions that maximize their interests given what they believe. Conversational moves should be calculated via an estimation of best return given what other participants say. This indicates that strategic conversations are games, and that game-theoretic analyses are a natural place to pursue the analysis of conversation. Consider a debate between two political candidates. Each candidate has a certain number of points she wants to convey to the audience; each wants to promote her own position at the expense of the other's. Such debates are typically 0 sum games. Typically only one agent can win, though there may also be draws. Similar strategic reasoning about what one says is a staple of board room or faculty meetings, bargaining sessions, and even conversations with one's children. Such conversations are common. To explain "what is going on" in such a conversation, we need to appeal to the participants' discourse goals, which may depend on the goals of the other participants.

Grasping the general goals of conversationalists do not suffice, however, to determine the structure of a conversation. Since conversations should be the result of rational inference to the best means for achieving one's conversational goals given one's information about the discourse context, each *linguistic move* in a conversation should ideally be related to general conversational goals. For cooperative conversations, we need to describe the linguistic reflection of the reasoning about a common decision problem, and this means we need to talk about the way clauses in a text rhetorically relate to each other in a way that has become familiar from theories of discourse structure like RST or SDRT [28, 1] and how this structure implements a plan to solve the decision problem. The interaction between goals and particular moves is important for understanding monologue as well, as one can ask what "problem" the author was trying to solve in a particular passage; there is a close correspondence between a coherent text's discourse structure and the text's "goal". We aim to tell a similar story for conversations in not necessarily cooperative settings. Given certain general conversational goals

for our conversational participants, we want to track how particular discourse moves detailed in a theory like SDRT takes a dialogue agent towards her conversational goals or thwarts them.

To get a better idea of the structure of conversations in strategic settings, we start from two intuitions. One is that many strategic conversations have a determinate outcome. One dialogue agent can "win" if she can play certain conversational moves; and if she does not, she loses. Another intuition is that in many conversations some conversational strategies, and some winning conditions or conversational goals, are more complex, and more difficult to achieve, than others. Understanding and categorizing such winning conditions and their strategies are an important part of understanding the large scale structure of conversations. In addition, they also determine whether a conversational agent has won the strategic conversational game, which is one important communicative effect of the conversation. But how can we measure or compare such strategies? This paper systematizes these intuitions and offers an answer to our question.

To strengthen these intuitions, here are some examples of conversations with their intuitive winning conditions.

Example 1. Suppose a candidate, Candidate A, has a joint interview with another competing candidate, Candidate B, for an academic position. Suppose Candidate A has proved an important theorem and she knows that during the interview if she can mention this, she will have "won" the interview by getting the job over the smarter candidate B, as long as she can mention this fact, no matter at what point of the meeting she says so. This is her winning condition.

Example 2. In the following example from [40], a prosecutor wants Bronston to say whether he had a bank account in Switzerland or not; Bronston does not want to make such an admission. His winning condition is to not answer the question directly, but only to implicate an answer that he doesn't have a bank account. He does not want to commit either to having or to not having a bank account.

(2)
a. Prosecutor: Do you have any bank accounts in Swiss banks, Mr. Bronston?

b. Bronston: No, sir.

c. Prosecutor: Have you ever?

d. Bronston: The company had an account there for about six months, in Zurich.

A non-courtroom variant of (2) is (1). The background is that Janet and Justin are a couple, Justin is the jealous type, and Valentino is Janet's former boyfriend.[1]

(1) a. Justin: Have you been seeing Valentino this past week?

b. Janet: Valentino has mononucleosis.

Janet's response implicates that she hasn't seen Valentino, whereas in fact though Valentino has mononucleosis she has been seeing him.

Example 3. Consider a *voire dire* examination in a medical malpractice suit from [27] where the plaintiff lawyer (LP) has as a goal to return repeatedly to the topic about the division of a nerve during a surgery. This goal has a further objective of getting the witness (D) to characterize the surgical operation as incompetent and mishandled. Repeatedly coming back to the topic can wear D down as actually happened in the case we cite.

(3) a. LP: And also, he put an electrical signal on that nerve, and it was dead. It didnt do anything down in the hand, it didnt make the hand twitch?

b. D: Correct.

c. LP: And we know in addition to that, that Dr. Tzeng tore apart this medial antebrachial cutaneous nerve?

d. D: Correct.

e. LD: Objection.

f. THE COURT: Overruled.

g. D: Correct. There was a division of that nerve. Im not sure I would say tore apart would be the word that I would use.

h. LP: Oh, there you go. Youre getting a hint from your lawyer over here, so do you want to retract what youre saying?

The defendant was resisting this line of attack relatively well, but then made an error by agreeing to LP's loaded question.

Example 4. During the Dan Quayle-Lloyd Bentsen Vice-Presidential debate of 1988, Quayle was repeatedly questioned about his experience and his qualifications to be President. Quayle attempted to compare his experience to the young John Kennedy's to answer these questions; his winning

[1]Thanks to Chris Potts and Matthew Stone for this example.

condition was probably to suggest with this comparison that like Kennedy he was a worthy Presidential candidate. Part of his goal too was to have this comparison pass without criticism (perhaps because he couldn't defend it adequately), and so it was indirect. However, Bentsen made a discourse move that Quayle didn't anticipate.

(4)
a. Quayle: ... the question you're asking is, "What kind of qualifications does Dan Quayle have to be president," [...] I have as much experience in the Congress as Jack Kennedy did when he sought the presidency.

b. Bensten: Senator, I served with Jack Kennedy. I knew Jack Kennedy. Jack Kennedy was a friend of mine. Senator, you're no Jack Kennedy.

c. Quayle: That was unfair, sir. Unfair.

d. Bentsen: You brought up Kennedy, I didn't.

Bentsen's surprise move successfully attacked Quayle's strategy to establish a comparison between himself and John Kennedy. Quayle had no effective defense and lost the debate handily.

Example 5. Allegedly, the physicist and Nobel laureate Richard Feynmann decided the topics of his next lecture in advance and prepared for it for over 8 hours. However, when he entered the class he would start off with: "So what shall we discuss today?" But he would always have a strategy to steer the conversation to the topics he had prepared for, whatever his students, who always wanted to stump him (and so had opposing interests to Feynmann's), would answer. Feynmann's winning condition was eventually to get to his prepared topic and stick to it for the remainder of the lecture.

Our examples so far have described or been excerpted from actual finite conversations that are relatively circumscribed. But conversations can occur over a much longer period, say over an entire Presidential campaign as in our next example. Nevertheless, they are still *linguistic conversations*.

Example 6. Recall President Clinton's adage "it's the economy stupid." What Clinton meant is that he should keep the conversation focused on questions concerning the economy in the extended debate between his Democratic team and the opposing Republican one during the 1992 Presidential campaign. As long as Clinton was able to bring the debate repeatedly back to a discussion of the economy, he achieved his winning condition.

We claim the following are important features of strategic conversations (and perhaps of conversations generally).
(I) People have conversations for purposes. Their conversations are successful when they achieve those objectives. Crucially, some, perhaps most, of these objectives involve commitments to contents, which are the conventional meanings and contextually derived implicatures of the utterances of the conversation. The commitments essential to realizing a participant's conversational goals may be those the participant makes herself, but some may be commitments made by other conversational participants.
(II) In principle, conversational players have no limits on the length of their intervention, though they are finite. In practice exogenous time limits may be imposed.
(III) Players can in principle "say anything" during their conversational turn, though what they say may very well affect whether their conversation is successful or not.
(IV) While conversations are finite, they may have no designated "last turns;" conversational agents cannot in general foresee who will "have the last word." Hence, people strategize in conversations even when they cannot anticipate when the conversation will end, what possible states might arise, or what utterances their opponent will consider.

In order to turn features (I-IV) into a model, we need three things: (i) an appropriate vocabulary of conversational moves for building sequences of message exchanges between players, (ii) goals or winning conditions for conversational players, (iii) a way of modeling the epistemic limitations that players cannot in general foresee the last move of a conversation. Infinitary games like Banach Mazur (BM) games [32, 19, 23] furnish a good point of departure, as they reflect some features of (I-IV). For simplicity, we will mostly restrict our attention to two-player win-lose games, allowing us to concentrate on basic conceptual points, though in section 4 we briefly consider extensions. Our theory distinguishes between conversations in virtue of their winning conditions, which are global properties of the entire conversation. Different winning conditions require different strategies for achieving them giving rise to different linguistic realizations. We give a precise criterion for the existence of these strategies and a formal model of winning conditions, enabling us to compare different conversational goals and their winning strategies. No extant framework examines the structure of conversations in general and their game theoretic structure in such a precise way.

Our paper is organized as follows. Section 2 introduces the basic points of our model in more detail and considers related work; section 3 gives back-

ground on infinitary games and introduces our theory of message exchange games. Section 4 develops a typology of conversations via their winning conditions. We investigate constraints on winning conditions that are intrinsic and normatively necessary for winning conversations most of the time, such as consistency and discourse coherence. These constraint render conversations more complex. Section 5 concludes our paper with some pointers to future work.

2 Conversations as infinite games

As stated in the introduction, we think of conversations as games in which the players are trying to achieve a certain end, that the conversation go in a particular way. These games involve a set of sequences of conversational moves and a characterization of winning conditions for players of the game. We now delve deeper into the structure of conversational games. What do they concretely involve? What are the 'moves' of the players, what are their 'strategies' and so on? What are their winning conditions? And how can we model conversational goals in a formal setting?

2.1 Signaling games, Grice and opposing preferences

Game theory has had several applications in pragmatics [25, 33, 34, 35, 7, 14, 16, 43, 44]. Much of this literature uses the notion of a signaling game, which is a sequential (dynamic) game in which one player with a knowledge of the actual state sends a signal and the other player, who has no knowledge of the state, chooses an action, usually an interpretation of the signal. The standard set up supposes that both players have common knowledge of each other's preference profiles as well as their own over a set of commonly known set of possible states, actions and signals. The economics literature contains a detailed examination of signaling games, [41, 11, 13, 37], to name just a few important papers in this area.

Although they have proved useful for many issues, signaling games do not offer a straightforward way to encode the principles we outlined in the introduction, especially for strategic contexts. As a consequence, the model we propose in this paper differs from signaling games in many aspects. However, it does not contradict the predictions of signaling models but rather provides a natural and convenient way of addressing situations that are not transparently expressible as a signaling game's context. We now explain why the strategic contexts considered in this article fall into this category.

A game requires, in order to be a reasonable candidate for modeling non-cooperative contexts, that its structure encodes the players' divergent preferences. As emphasized earlier, the most intuitive and simplest way of doing that is to assume a 0 sum game. Signaling games however predict that no communication happens in such games and signals convey the information that all states are equally probable: a consequence of the results in [11] is that in equilibrium[2] the sending of any message has no effect on the receiver decision.

An immediate corollary is that assuming that the sender has the possibility of (costlessly) not sending a message and that the sending of any message has at least an infinitesimal cost, ϵ, makes it optimal for the sender to not send anything. This leads to obvious, unintuitive and irrational consequences. Hence, the most straightforward way of setting up non-co-operativity makes communication of any kind is impossible in a signaling game. This means that non-co-operativity of the sort we are interested here should not translate as 0-sum utilities in a signaling model.

Still, there is, between perfectly aligned utilities and 0-sum games, a space of games with partially aligned utilities which could encode some lack of co-operativity into the context while still allowing for communication to take place. Notice that yielding the right equilibrium is not the only demand on the game structure: a precise justification of the chosen utility profile is also needed. In order to use signaling games as part of a general theory of meaning, one has to make clear how to construct the game-context, which includes providing an interpretation of the game's ingredients (types and actions) and explaining why the utility profiles fits the situation to be modeled. [15], for instance, associates in a principled way an *interpretation game* to a given utterance. Interpretation games form a subclass of signaling models that assume a specific class of sender types, actions and preferences. They intend to encode a "canonical context" for an utterance, in which relevant conversational implicatures may be drawn. In interpretation games, the full game structure is determined by the set of sender types: there is a bijection between the set of receiver actions and the set of sender types, and the utility profile is such that both the receiver and sender get rewarded if they coordinate on the sender actual type, and do not gain anything otherwise.

[2]Assuming bounded rationality of conversational agents may restore an effect to messages: for instance the Iterative Best-Response model in [15] allows a level 2 sender to misdirect a less sophisticated level 1 receiver. However, we are convinced that the conversational examples presented in this article are compatible with a common belief in rationality and require an analysis making such an assumption.

Such a setting is very intuitive and interestingly does not seem to require further precision on what exactly it means for the receiver to take the action a_t associated with receiver type t and why such an action should indeed maximize the receiver payoff if t is the sender's actual type. Of course, one can still wonder whether performing a_t means, for instance, that the receiver believes that the actual state is t (let us call this option 1), or that the receiver interprets the sender's message as a commitment to the actual state being t (option 2), or even that she herself commits to the sender committing to the actual state being t (option 3). But, despite this ambiguity between, in principle, distinct ways of understanding action a_t, there is, in interpretation games, no necessity to settle the question, because, for **Gricean agents**, the different options collapse. A Gricean sender should intend to commit to what he believes is true (sincerity), co-operativity should make a Gricean receiver intend to interpret the sender commitment as what the sender intends to communicate, and belief in the sender's sincerity should make him believe that the sender believes in what he has committed to. Hence options 1 and 2 collapse. Finally the receiver's sincerity and co-operativity ensure that (at least if the sender's needs her to), she acknowledges that the senders' commits to what she interprets the sender to commit to (namely, that t is the actual sender's state). Hence option 2 implies option 3, and the other way around follows from the assumption of the receiver's sincerity. Therefore, the games structure as it stands seems to offer a perfectly adequate level of abstraction.

But things become much more intricate as soon as one is considering potentially non-Gricean players, and this makes the task of understanding and providing justification for a (partially) unaligned utility profile much more involved. It depends on what one takes actions and types to represent. Recall (2) and imagine, for the sake of argument, that we want to model Bronston's answer with a signaling game involving two sender types: t_{BANK} and $t_{\neg\text{BANK}}$, two corresponding interpretative (or acknowledging) actions a_{BANK} and $a_{\neg\text{BANK}}$, and three possible messages, $m_{\text{YES}}, m_{\text{NO}}$ and m_{COMPANY}. These messages are respectively true in the sets of states $\{t_{\text{BANK}}\}$, $\{t_{\neg\text{BANK}}\}$ and $\{t_{\text{BANK}}, t_{\neg\text{BANK}}\}$. Assume also that we want to accommodate a fear of perjury on Bronston's part into the game context. Consider first that performing action $a_{\neg\text{BANK}}$ means for the receiver to update his belief to include that $t_{\neg\text{BANK}}$ is the actual sender's type, or at least, to subsequently acts as if it were the case. Under such an interpretation, if the sender, Bronston, sends m_{NO} and the prosecutor takes in return action $a_{\neg\text{BANK}}$, should Bronston fear being charged with perjury? Intuitively no, because such an attack would indicate an inconsistent belief of the prosecutor that $t_{\neg\text{BANK}}$

holds (because the action he took is interpreted as such) and does not hold at the same time.[3] Then again, if actions are to be interpreted at the level of public commitments (say, using option 3 of the previous paragraph), taking action $a_{\neg \text{BANK}}$ after receiving m_{NO} commits the prosecutor to the proposition that Bronston is committed to $t_{\neg \text{BANK}}$. This does not imply that the prosecutor believes (or commits to) the latter state to be actual. Hence, if the prosecutor takes this action, he is susceptible to attack Bronston for perjury by committing to Bronston's actual type being in fact t_{BANK} and not $t_{\neg \text{BANK}}$. Bronston's payoff in that case should depend on whether the prosecutor will indeed charge him with perjury and how bad the consequences will be. If the prosecutor has solid arguments to show that t_{BANK} is the actual state, then Bronston should fear such a continuation and have a very low payoff for the triple $\langle t_{\text{BANK}}, m_{\text{NO}}, a_{\neg \text{BANK}} \rangle$. Bronston should be better off with the triple $\langle t_{\text{BANK}}, m_{\text{COMPANY}}, a_{\neg \text{BANK}} \rangle$ as, in that case, he disposes of a way to defend himself against a charge of perjury, which consists in committing that he never committed to $t_{\neg \text{BANK}}$. Put another way, he can argue that the prosecutor with $a_{\neg \text{BANK}}$, committed to something false –namely that Bronston committed to $t_{\neg \text{BANK}}$. Notice that such a defense is not a very good option when Bronston sends m_{NO}, because in that case, it requires Bronston to say something false[4]. Notice also, that resorting to this defense should bear a non-negligible cost: although he avoids perjury, Bronston can be asked to answer why he did not respond to the the prosecutor question in the first place and/or why he did not immediately correct the prosecutor after the latter performed $a_{\neg \text{BANK}}$.

These considerations illustrate two things: first, if not all agents conform to Gricean maxims, different choices in the way to interpret actions and types yield different games context with different predictions. Hence it becomes primordial to make precise what the exact set of actions is and what they represent —something which may vary according to the nature of the player's objectives (commitments, beliefs, both, something else, ...). Second, the payoffs of the sender and receiver may depend on subsequent actions, which requires that the possible outcomes of the signaling games encode all possible relevant continuations of the conversation. None of this

[3] We assume here that the prosecutor has an interest to charge Bronston with perjury only if he believes that Bronston actually performed perjury. One can relax this assumption, but that would mean that the prosecutor's beliefs are irrelevant to his subsequent moves and that the commitments-related interpretation of actions should be considered here.

[4] and following the logical model of commitment that one adopts, it can even make him inconsistent. See [45] for a discussion.

is self-evident (and we will examine some reason for why it is so) and makes a systematic construction of a game context much more difficult than in the cooperative case. These difficulties echoes the close correspondence between a general formalization of Gricean principles and that of games with shared interests that **(author?)** [2] establishes. This does not entail that in strategic settings, Gricean principles do not ever apply, but the result does establish that a player i shouldn't count on Gricean principles as operative; in general i can't assume that players are maximizing quality, quantity or relevance (to i's own conversational ends).

Furthermore, there is a crucial difference between being a non-Gricean speaker, and admitting to being so. A player's conversational objectives are very likely to include not making such an admission. Conversations can thus involve sorts of hide-and-seek games where agents try to expose the "bad" behavior of their opponent while making themselves look good. In other words, bad (typically, non-Gricean) behavior is licensed as long as it remains hidden or deniable. In 2, if the prosecutor rejects Bronston's indirect answer as being insufficiently informative, he commits at the same time that Bronston's co-operativity is, at least, subject to caution. If Bronston then admits having had a swiss bank account, he justifies this cautious attitude, and commits to the proposition that he was not cooperative in giving the indirect answer. In such a context, the prosecutor intuitively should claim that Bronston is being non-cooperative but Bronston should try to avoid admitting that he is. In analogous contexts occurring outside of the courtroom, it might be rational for an interrogator who cares for his reputation or his interlocutor's friendship to prefer a misleading answer over formulating a public accusation of non co-operativity that he cannot prove.

Signaling games being one-shot games basically leaves the modeler with two choices: either the game's payoffs are locally determined, and then the game, by choice of design, does not evaluate the long term impact of a triple ⟨type,message,action⟩, $\langle t, m, a\rangle$, or the payoffs are global, but then must somehow be computed taking every possible continuation of $\langle m, a\rangle$ into account (if this is possible at all, which is discussed more at length in 2.2). There is also an afferent problematic asymmetry of the sender and receiver in such games. While, in signaling games, the sender, sending a message, may reveal information about his type and associated preference profile that the receiver may seize on, the latter does not reveal anything that is not already common belief when he chooses an action. There is anyway no subsequent move of the sender to reward or punish such a revelation by the receiver.

Again, this is problematic because it blurs the frontier between public and private information which an accurate analysis of strategic settings

requires to be clearly marked. While types, preferences and beliefs are intuitively private, messages are intuitively public. The nature of receiver's actions however, as we have seen, is not clear: belief update or interpretative actions are private actions whereas acknowledgments, responses are not. Moreover neither option is plainly satisfactory (for a full account of strategic context): if actions are private updates, then they are not sufficient on their own to constrain subsequent conversational moves, unless the model is supplemented with a theoretical link between private and public attitudes (once again, this is precisely what a formalization of the Gricean maxims provides **for cooperative settings only**). If actions are public, then the rational receiver taking a particular action reveals, at the same time, having a certain belief about the sender's state, namely one that makes the chosen action rational. Even if we assume this belief to be common to the sender and the receiver, it does not mean that either of the two is comfortable with making this belief into a public commitment. There can be numerous reasons for avoiding such an admission: for instance, a third party monitoring the conversation, and/or politeness constraints preventing the players to put their cards on the table. Imagine two agents holding each other in very low regard, with a common belief thereof: it seems reasonable to imagine keeping the conversation polite and cordial as one of their objective, restricting without voiding the moves they might use toward a second objective of defending a position at the expense of the other. They should, in particular never explicitly commit to what they think of each other despite being fully aware of it on both sides. Yet, due to signaling games' asymmetry, it seems much easier to implement these constraints on the sender's side: to represent a sender who wants to avoid a commitment to having the preferences associated with a given type t_{BAD}, it suffices to model the bad consequences of such a commitment, simply assuming an action $a_{t_{\text{BAD}}}$ optimal for the receiver iff t_{BAD} is the sender's actual state, and which, in that case, comes with a dramatically negative payoff for the sender. Doing the same on the receiver's side, can, at best, only be achieved by refining the set of sender types so as to distinguish between sender types according to (at least some of) the subsequent conversational moves that the sender will perform. But such a trick essentially amounts to consider sequential games with more than a single turn, which we advocate in the next sections.

We conclude this discussion with a more formal analysis of the above points and an example: how does *deception* in the sense that we encountered in our examples translate in signaling terms? We can define it this way: a sender S of type t, deceives a receiver R if he can induce an action a' from the receiver which is suboptimal for him given $\langle t, m\rangle$, *i.e.* such that there

exists another action a verifying $u_R(t, m, a) > u_R(t, m, a')$ [where u_R and u_S are the utility functions for the receiver and the sender respectively]. Let, for any pair $\langle t, m\rangle$ of type and message, $a^*_{t,m}$ denote an optimal action for a receiver knowing that t is the actual state and m was sent. Following our notion of deception, we say that the type t_{BAD} has an *interest to deceive* using message m iff there is a type t_{GOOD} such that we have:

$$u_R(t_{\text{GOOD}}, m, a^*_{t_{\text{BAD}},m}) < u_R(t_{\text{GOOD}}, m, a^*_{t_{\text{GOOD}},m}) \text{ and}$$
$$u_S(t_{\text{BAD}}, m, a^*_{t_{\text{BAD}},m}) < u_S(t_{\text{BAD}}, m, a^*_{t_{\text{GOOD}},m})$$

Let T denote the set of sender types and T_{GOOD} denote the set of "good" sender types, *i.e.* the subset of the set of sender type for which the receiver has better payoff using $a^*_{t_{\text{GOOD}},m}$ than $a^*_{t_{\text{BAD}},m}$ upon reception of m:

$$T_{\text{GOOD}} = \{t \mid u_R(t, m, a^*_{t_{\text{BAD}},m}) < u_R(t, m, a^*_{t_{\text{GOOD}},m})\}.$$

Let $T_{\text{BAD}} = T \setminus T_{\text{GOOD}}$. We shall denote a typical element of T_{GOOD} as t_{GOOD} and that of T_{BAD} as t_{BAD}. Using the above notation, we have the following:

Proposition 1 *Whenever a rational receiver R is deceived by m,* i.e. *he takes $a^*_{t_{\text{GOOD}},m}$ with non zero probability after receiving m from a sender S of type t_{BAD}, he must believe after reception of m that S being of a type in T_{GOOD} is at least $\delta_{\text{GOOD}}/\delta_{t_{\text{BAD}}}$ as likely as S being of type t_{BAD}, where $\delta_{t_{\text{BAD}}}$ is the gain in payoff for the receiver choosing $a^*_{t_{\text{BAD}},m}$ rather than $a^*_{t_{\text{GOOD}},m}$ in state t_{BAD}, and δ_{GOOD} is the maximal gain in payoff for the receiver choosing $a^*_{t_{\text{GOOD}},m}$ over $a^*_{t_{\text{BAD}},m}$, in any state of T_{GOOD}.*

To preserve the flow of the text, we provide the proof of Proposition 1 in the Appendix. As an immediate corollary, it follows that in any perfect bayesian equilibrium with a pure sender strategy sending m in state t_{BAD}, if R is deceived then the prior probability of being of a "good" state in which S sends m must be at least $\delta_{\text{GOOD}}/\delta_{\text{BAD}}$ as likely as the prior probability of S being of state t_{BAD}.

What this fact shows is that, in a signaling game, the only basis for a receiver to ever accept a deceptive move (*e.g.* a misleading answer) is that he judges it more likely (modulo the ratio $\delta_{\text{GOOD}}/\delta_{\text{BAD}}$ which quantifies the "badness" of the deception) that his opponent is of a "good" type, never that he lacks an argument to confront him, or has reasons to avoid confrontation. Yet we are convinced that the latter are equally good reasons to accept "dodging" moves. Consider as an example the following conversation:

(2) a. A: Are you available on Thursday afternoon? I need help moving in.

b. B: I have a very important meeting on Thursday.

It seems intuitive to assume that B might be of two distinct types: the "free on Thursday afternoon" type t_{F}, and the "not free on Thursday afternoon" type $t_{\neg\text{F}}$. Assume further that B's meeting is scheduled early Thursday morning, so that despite telling the truth, B is in fact of type t_{F} (were B willing to, he would still be able to help A in the afternoon). A's best interest in that case involves intuitively to have B commit to being of type t_{F} as it makes it socially difficult for B to refuse his help (and unless A has, and is willing to use, another way to pressure B, this is likely to be the best he can achieve by talking). Assume finally that A, for some reason, has an incentive to not trust B and think that it is much more likely that is available than not (because for instance, they work at the same company and B has a reputation of being very lazy). Let us have a look at some of the possible responses A can make:

c_1 A can take B's answer has a no with "A: OK, too bad".

c_2 A can request a direct answer "A: So you are not available on Thursday afternoon?"

c_3 Depending on his relation with B, A can try other, more gentle or less direct denunciations of B's misdirection, for instance we can imagine something in the spirit of "A: Oh, I forgot about that! Those morning meetings can be exhausting, I take it that you'll be too tired in the afternoon then?"

Naturally, the best option for A depends on his familiarity with B, but the point still, is that the payoff of continuation c_2 highly depends on B's reaction. In many cases, A might judge c_2 too abrupt, and not really helpful in the case of B opting for an explicit *no* answer in return. A might thus think that he has little to gain using c_2, and B, counting on that, even if he knows that A believes him to be more likely available than not, will still try to misdirect, expecting A to opt for c_1, or at worse for something like c_3 which, unlike c_2, saves his face.

Indeed, if A opts for c_2, B might get annoyed and react badly with a reply like: "No, as I just said, I have an important meeting."[5] Arguably this

[5]Notice also that, interestingly, even if B explicitly lies about his availability with such an answer, he remains only implicitly committed that it is the meeting that makes him

should yield a bad payoff for both A, who has not fulfilled his objective,[6] and B, who has been compelled to lie. But importantly, it is also possible that B alternatively chooses to drop his attempt at misdirection by *e.g.*, saying "Well, actually my meeting is in the morning, so I guess I'm free after all." In that case, c_2 should yield high payoff for A and low payoff for B.

In summary, these considerations show that the rationality of B's attempt at a misleading answer is much more dependent on the probability of the different continuations of c_2 than on A's prior belief about B's actual type. Is it possible to account for this using a signaling game? From what we have shown, it is, but only at the cost of an "odd", non transparent payoff structure, and/or a rather complicated type space: let `meeting` denote B's answer to the question. Let u_A and u_B be the utility functions of A and B respectively. As B's objective is to avoid committing to t_{F}, we should have $u_B(t_{\text{F}}, \texttt{meeting}, c_2) < u_B(t_{\text{F}}, \texttt{meeting}, c_1)$. The real difficulty is to set the payoff for c_2. If we let $u_A(t_{\text{F}}, \texttt{meeting}, c_2) > u_A(t_{\text{F}}, \texttt{meeting}, c_1)$ then, given A's prior, we are in a case of application of 1 and B is not rational in trying to misdirect. If we let $u_A(t_{\text{F}}, \texttt{meeting}, c_2) < u_A(t_{\text{F}}, \texttt{meeting}, c_1)$, then an important part of the reasoning becomes hidden in the payoffs, and the games becomes "artificially" cooperative: any explicit encoding of the opposing interests at stake is left behind.

Again, a solution to this dilemma is to refine the set of types distinguishing for instance between B's of type t_{F} that reacts aggressively to c_2 with the full lie and those that would rather drop their misdirection attempt. But this means that we must condition sender strategies to their type, and more generally this boils down to considering another class of games that involve more than one send and response.

Other models like that of [18] exist that do not use signaling games. However, they also have difficulties in expressing the sort of constraints we have developed above. Signaling games and persuasion games both still take a broadly Gricean view of communication: conversations are essentially information gathering or exchange activities; agents exchange messages for the purpose of affecting the beliefs of the other partner. This is precisely, however, what is in doubt in many conversations. In many conversational settings and in all of our examples, agents converse not in the hope persuading their opponents, but rather to impress or persuade others, and perhaps

unavailable. So he can still drop this commitment at the cost of admitting that he was incoherent or not responsive. Hence, even if A, for some reason, is willing to confront B for lying to him, and has formal evidence that the meeting is indeed in the morning, doing so still requires a lot of efforts on his part.

[6] unless of course A is ready to accuse B of lying to him.

themselves. Just as Grice captures important aspects of some but not all conversations, people do try sometimes to persuade or to exchange information, but this is not a general framework for all conversations.

We need a different game-theoretic model of conversation. Our players interact with each other and exchange messages that convey objective, public commitments. For instance, D in (3d) commits to Dr. Tzeng's having "torn apart" the nerve by agreeing with LP's description in (3c). D then tries to go back on that commitment in (3g). D may or may not believe this commitment. But if he agrees with (3c), then he is committed to its content, and he can be attacked on the basis of that commitment or subsequent commitments. As the excerpts in our examples make evident, conversationalists often pay careful attention to the commitments of others, not only to explicit commitments but also to their implicatures. For example, in (4) Bentsen seizes on a weak or possible implicature of Quayle's commitments, that he is comparable in Presidential stature to JFK, and attacks Quayle for it. In our examples, the moves players make to defend their commitments or to attack those of an opponent exploit the conventional meanings and even the implicatures that messages have. So our model must enable us to fix the meanings of players' moves to their conventional meaning, regardless of what the players themselves believe.

But why do conversationalists make the commitments they do, if they don't do it to persuade their interlocutors or to send a signal that their interlocutors will find credible? Players make the commitments they do, for the purpose of convincing or influencing a third party, which we call *the Jury*. The Jury is for us an abstract role that can be satisfied in diverse ways. In examples (2) and (3), it's the jury of the court; in examples (4) and (6), it was the American electorate. Sometimes the third party may be one of the players, as in example (1). The Jury does not as such participate in the conversational game but is rather a scoring function for the game. Players choose their conversational objectives based on what they believe they can defend against their opponents and that will find favor in some way with the Jury. A player attempts to achieve her conversational objective, while her opponent tries to thwart her. The Jury is an unbiased, rational and competent user of the language of the players and judges on that basis whether a given discourse move or a sequence of moves contributes toward the realization of the conversational objectives of a player or not. We make the simplifying assumption for most of the paper that the Jury can only be convinced by one player.

2.2 Why infinite games?

We believe that humans must act as though conversational games were unbounded. If conversations have definite last moves and our players have opposing interests, even the presence of the Jury will not explain why our agents converse in the way they do.

Consider example 2 again, or its non courtroom variant (1). Janet is presented with a Hobbsian choice. Ideally, she would prefer not to answer the question at all or simply lie. To not answer the question or to lie would be rational and that is what Janet should do, if she were playing a one shot game with no further interaction with Justin (this is akin to the defect move in the Prisoner's Dilemma). As conversations, however, have continuations, many people have the intuition that a refusal to answer will make Janet fare worse in subsequent exchanges. Janet cooperates with her interlocutor in the minimal sense of providing a response to the question, what [2] call *rhetorical co-operativity*, because of reputation effects. If Janet does not cooperate by responding to Justin, she risks receiving uncooperative treatment if in the future she asks a question or make some demands of him. This is a form of the "tit for tat" view of [6].

Nevertheless, the reputation argument has its problems. Conversations are not just a matter of retaliating in kind with respect to your interlocutor's actions; there are other constraints operative on conversations. Further, if a conversation is just a finite sequence of one shot games, what holds for a one shot game holds throughout a conversation. Backward induction over such a finite sequence would lead Justin to the conclusion that he should not bother to ask his first question because it is in Janet's to defect at the earliest possible opportunity. If there is a foreseeable last move for one of the players i, then she will play to her advantage and defect on the last move, if her opponent has gone along in the discussion up to that point. The opponent seeing this will reason by backwards induction to defect at the earliest possible moment, and the speaker will recognize this fact. The prediction is that given a foreseeable last move, no message exchange should occur.

If the conversational game is assumed to be infinite, however, the formal argument for the rationality of defection over sequences of exchanges in cases where conversationalists have opposing interests disappears. The argument from backward induction fails because there is no last move from which to begin the induction. However, there is still some explaining to do. A simple "tit for tat" model does not explain why interlocutors cooperate with each other rhetorically, *even if* their roles *vis a vis* their interlocutors

are never reversed, even if Janet or Bronston never make any demands of their interlocutors.

In our model, the Jury can force rhetorical co-operativity. A defection will hurt player i if the Jury can infer that i's is defecting because a rhetorically cooperative move would reveal a reason for them not to be persuaded by her. This is also a feature of the model of [17, 18] but their model is more restrictive. In their model, the Jury only interacts with one sender who must persuade the Jury to accept or reject a message. In addition, the sender of a message is restricted in her choice of messages she can send in a given state, and so the Jury can draw more secure inferences from messages she does not send. Since she can only send certain messages in certain states (e.g., Bronston might be able to say he did not have an account only in a state where he truly does not), a failure to send a message or to respond to a question where the message is directly requested and would be in the player's interest to send could well indicate that the player is not in the state where such a message is permitted.

We have made no such assumptions about messages, however, because we do not think that messages are tied to states in such a simple way. One can say anything, regardless of the state of the world in a conversation. So the reasoning from signals and strategies to the persuasiveness of a player is much more uncertain for the Jury in our games. In addition, while Player i needs to convince the Jury that she has achieved her conversational goals, her goals crucially involve her opponent and hence are more complex than simply getting the Jury to accept the content of a particular message. Player i could simply refuse to cooperate with her opponent, because she has a general strategy of not revealing information to her opponents. Or she could provide a reasonable defense for why she is not cooperating. In either case, it falls on *the opponent* to make the case to the Jury that player i's lack of rhetorical co-operativity provides a reason to deny i victory. If attacked, player i can reply to the opponent, defending her lack of co-operativity, and then the opponent must press the issue.

The following excerpt from a press conference by Senator Coleman's spokesman Sheehan brings out these features of our model. Senator Coleman was running for reelection as a senator from Minnesota in the 2008 US election (thanks again to Chris Potts for this example):

(3)

a. Reporter: On a different subject is there a reason that the Senator won't say whether or not someone else bought some suits for him?

b. Sheehan: Rachel, the Senator has reported every gift he has ever received.

c. Reporter: That wasn't my question, Cullen.

d. Sheehan: The Senator has reported every gift he has ever received. We are not going to respond to unnamed sources on a blog.

e. Reporter: So Senator Coleman's friend has not bought these suits for him? Is that correct?

f. Sheehan: The Senator has reported every gift he has ever received. (Sheehan continues to repeat "The Senator has reported every gift he has ever received" seven more times in two minutes to every follow up question by the reporter corps. `http://www.youtube.com/watch?v=VySnpLoaUrI`)

Sheehan, like Bronston, is seeking to avoid committing to an answer to a question. Sheehan's (3b) in response to the reporter's first question could be interpreted as an indirect answer, an answer that implicates a direct answer; the senator did not comment on the question concerning whether he had received the gift of suits because he felt he had already said everything he had to say about the matter. But in (3) the reporter does not accept this rather indirect answer; she says that Sheehan's response was not an answer to her question. In effect, she wants a direct answer to the question concerning the suits. Sheehan then explains why in 3d he will not answer the question. The reporter then presses the issue, and Sheehan becomes rhetorically uncooperative for the rest of the exchange, repeating the same thing. At this point, the Jury will begin to reflect on Sheehan's strategy: is he being rhetorically uncooperative because he has something to hide? His earlier explanation for his defection from rhetorical co-operativity becomes lost, and it becomes more and more plausible that Sheehan will not answer the question because the true answer is damning to his interests. To win given a defection from rhetorical co-operativity, Sheehan has to have a reply for every attack; if the opponent eventually introduces an attack for which he does not have a convincing reply (e.g., he simply repeats himself or simply stops talking), the opponent will win.

The need to justify uncooperative moves or defection generalizes. In most strategic situations, in order to win, 0 must engage with questions

and remarks of her opponent(s); she must show that her opponent cannot attack her position in such a way that a rational unbiased bystander would find plausible. For any discourse move, we can imagine a potential infinity of attacks, defenses and counter-attacks. In successful play, a player has to be able to defend a move m against attacks; she may have to defend her defense of m against attacks and so forth. This is a general necessary victory condition for 0. Let $attack(n, m)$ hold if move m attacks move n; commitments or types of discourse moves that generalize over more specific discourse moves[7] that are used to defend or attack commitments:

Observation 1 (NEC) *A play is winning for* 0 *only if for all moves* n *of* 0 *and for all moves* m *of* 1, $attack(n, m) \rightarrow \exists k(move(0, k) \wedge attack(m, k))$

Conversely for 1, a sufficient condition for winning is the negation of Observation 1. Given (1) 0 wins only if she is prepared for the conversational game never to end and to rebut every attack by 1. It is this constraint that provides a second reason for assuming conversational games to be infinite and is a powerful reason for obeying rhetorical co-operativity.

NEC also has empirical consequences. In virtue of it, we can see why Quayle intuitively loses in example (4). Part of Quayle's winning condition was not to come under attack for his implicit comparison or at least to be able to rebut any attack on his move; that is NEC was also part of his winning condition. But given that he had no rejoinder to Bentsen's unanticipated move, he failed to comply with Observation 1 and so lost.

To Observation 1, we add another, motivated by example (3): to win, 0 should not simply repeat herself in the light of a distinct move by her opponent (at least not more than twice).

Observation 2 (NR) *A play is winning for* 0 *only if there is no move* k *by* 0 *such that* 0 *repeats* k *on* m *successive turns, for* $m \geq 3$*, regardless of what* 1*'s intervening contributions are.*

NR buttresses NEC 's support for rhetorical co-operativity.

While we could weaken the quantifiers in NEC and NR to something like *for most moves of 0*, we are rather interested in the general upshot of such constraints: to model winning play by 0, we need to model a conversation as a potentially unbounded sequence of discourse moves, in which she replies to every possible attack by her opponent. Moreover, at least some of the

[7] Examples of such moves are Answering a question, Explaining why a previous commitment is true, Elaborating on a previous commitment, Correcting a previous commitment, and so on—in fact, these correspond to the discourse relations of a discourse theory [1].

moves of player i must be related to prior moves of her opponent. It follows that it is always risky, and often just rationally unsound, to play a rhetorically uncooperative move like defection without further explanation that is optimal only if it is the last move in a finite game. Defection from rhetorical co-operativity is possible, but it must be explained or defended in any winning play convincing the Jury. A player who plays a rhetorically uncooperative move opens himself up to an attack that will lead to a defeat in the eyes of the Jury, as in (3). That is, relatively weak and uncontroversial assumptions about the beliefs and preferences of the Jury validate rhetorical co-operativity as a component of any winning play.

Even if in practice, conversations do not go on forever, players have to worry about continuations of conversations and thus should rationally act as if a conversation were 'potentially infinite'. In such situations, a theory of finite play does not apply and one has to resort to infinite plays. This is why it is necessary to adopt a framework of infinite games. The really interesting difference between strategic and fully cooperative contexts is that in strategic contexts, but usually not in cooperative ones, the players must behave as if the game is infinite, because they do not know who is going to have the "last word". In Gricean contexts, both players are trying to achieve a joint goal, and all things being equal, they want to achieve it efficiently. That efficiency is what limits the length of the game. Signaling games are ideal for capturing these limited (especially one-off) interactions. But as soon as strategic considerations are introduced, this efficiency goes out the window.

By moving to a framework with unbounded conversational sequences, [4, 5] show how games with unbounded cheap talk, games involving extended conversations with an infinite talk phase consisting of a pattern of revelations and agreements ending ultimately in an action, make possible equilibria for players that are not available in one shot or even sequences of revelations of bounded length. While we have adopted the simplest of payoff structures for our study, our examples show that unbounded conversational sequences allow players to win conversations that they otherwise couldn't. Had the reporters in 3 been limited to one question and one follow up, they could not have successfully attacked Sheehan in the way they did.

A final reason in favor of using infinitary games is, paradoxically, their simplicity. Given that we cannot impose any intrinsic limitations as to the length of conversations, a formalization of purely finite conversations is more complicated. In an infinitary framework, it is also straightforward to model finite conversations. Finite conversations are not just conversations that stop but crucially involve a point of mutual agreement that the players have

finished [39]. We represent a finite conversation then as one in which a finite sequence terminates with an agreement on a special "stop" symbol that is then repeated forever. More than that, initial prefixes of infinite sequences will play a very important role in the sequel. While our models of conversations will be infinite sequences, all that we ever make judgments on are finite prefixes of such conversations. We will have to evaluate the play of players and whether they have met or are meeting their objectives on such finite prefixes.

3 Message Exchange Games defined

We have established that conversations should be modeled as some sort of infinite game. In this section we define such games, which we call *message exchange games*, formally, using Banach Mazur games, a well-known sort of infinitary game, as a departure point and point of comparison. We then make some remarks about the expressive capacities of our new framework and examine how it addresses the problems we found with the signaling game framework.

Let V be a finite, non-empty set. We sometimes refer to V as the vocabulary. For any subset A of V, A^* is the set of finite strings over A and A^ω the set of countably infinite strings over A.

Definition 1 (BM game) *A* Banach-Mazur game (BM game)*, $BM(V^\omega, Win)$, consists of an infinite set of strings V^ω together with a winning condition $Win \subseteq V^\omega$.*

The game proceeds as follows. Player 0 first chooses a *non-empty finite* string $x_0 \in V^*$. Player 1 responds by choosing another non-empty finite string $x_1 \in V^*$. Player 0 moves next choosing another finite string x_2. This process repeats itself forever yielding a play p, an infinite sequence of alternating moves by 0 and 1. Let ρ be the set of all such plays. Define the flattening $flat$ of a play $p = (x_k)_{k\in\mathbb{N}}$ as the infinite sequence eventually designed by the two players: $flat(p) = x_0 \cdot x_1 \cdot x_2 \ldots \in V^\omega$. Player 0 wins the game if $flat(p) \in Win$. Player 1 wins otherwise. A strategy s_i for player i, is a function from the set of finite plays to the set of finite strings, V^*. A play $p = (x_k)_{k\in\mathbb{N}}$ of the game, is said to be consistent with the strategy s_i iff, for every integer k, $k \bmod 2 = i \Rightarrow x_k = s_i(x_0 \cdot x_1 \cdot \ldots \cdot x_{k-1})$. In other words, each move of player i is played according to s_i. A strategy s_i is said to be winning iff in every play consistent with s_i, player i wins.

BM games suggest a natural model for conversations: participants alternate turns in which they utter finite contributions. These contributions add to each other, and together form a conversation. This process potentially goes on indefinitely, or, at least strategic reasoning requires thinking of it that way. However, BM games "erase" the information of who said what in the following sense. For any string $x \in V^\omega$ let $flat^{-1}(x)$ be the set of all plays whose flattening is x. That is,

$$flat^{-1}(x) = \{p \mid flat(p) = x\}$$

Proposition 2 *Let $BM(V^\omega, Win)$ be a BM-Game. Then, for any play p and every play $p' \in flat^{-1}(flat(p))$, player i wins p iff player i wins p'.*

Given any infinite sequence s, any countably infinite set $turns \subseteq \mathbb{N}$ such that $min(turns) > 0$ yields a play in $flat^{-1}(s)$ and conversely: every element in $turns$ specifies a position in s which is the end of a player's move. A corollary of the above proposition is that we cannot define a winning condition that imposes for instance that Player 1 says something in particular, as long as she and 0 do not infinitely repeat the same single move. We formalize this observation as follows:

Corollary 1 *Let $y \in V^+$. Let Player 0 win a BM game if and only if Player 1 plays y somewhere in a play. That is, suppose the set of winning plays for Player 0 is $\{p = x_0x_1 \ldots \mid \exists i \geq 0, x_{2i+1} = y\}$ Then Player 0 does not have a winning strategy.*

Proof Consider the play $p = u^{2i}yu^\omega$ such that y is a suffix of neither u^{2i} nor u^ω. Note that p is winning for Player 0 because of the decomposition of p as $p = x_0x_1 \ldots$ where $x_0 = x_1 = x_{2i} = u$, $x_{2i+1} = y$. But consider another play $p' = x'_0x'_1 \ldots$ of x such that $x_0 = x_1 = x_{2i-1} = u$, $x_{2i} = uy$. p' is losing for Player 0. But, $p' \in flat^{-1}(flat(p))$. Hence by Proposition 2, Player 0 wins p iff she wins p' which is a contradiction. □

BM games have two limitations in the analysis of conversation. The first limitation of BM games is that they can model only zero sum games. The second is that they erase turn structure, which is important given our principle I. We need a more structured type of game given our principle I. A given conversationalist might have as a goal that her interlocutor *and only she* commits to a particular content, or answers a particular question, which BM games do not allow. Consider again (2). There's an important difference between Bronston's response to a question by the prosecutor and

the prosecutor's offering that information himself, and that difference can't be captured in BM games under all interesting scenarios. Discourse moves contain more information than the sentence itself. The discourse move that Bronston commits to a negative answer to (2a) provides more information than just the string *no sir* provides, and it is such moves that are of interest.

To remedy these limitations, we introduce Message Exchange (ME) games. An ME game involves a vocabulary of discourse moves V_i for each player i, $i \in \{0, 1\}$. If the vocabulary V is common for both the players then we let $V_i = V \times \{i\}$ to exemplify the turn structure (as in who played what). An ME game is an infinite game where the players 0 and 1 alternate by playing finite sequences of moves from V_0 and V_1. ME games include zero sum games as a special case:

Definition 2 (ME game) *A* Message Exchange game (ME game)*, $\mathcal{G}$, is a tuple $((V_0 \cup V_1)^\omega, Win_0, Win_1)$ with $Win_0, Win_1 \subseteq (V_0 \cup V_1)^\omega$.*

The game is structured so as to encode the information which player played what into the plays, and proceeds as follows: a turn by i is a non empty finite sequence of elements in V_i. The game is played similar to a BM game. The players play alternatively, with Player i, at turn j, picking a finite $n$on-empty sequence x_j of moves from her vocabulary V_i, which we will interpret as a finite (dynamic) conjunction of move formulas [we describe it in more details in Section 3.1]. A play p is thus an infinite sequence $p = x_0 x_1 \ldots \in (V_0 \cup V_1)^\omega$ with $x_j \in V^*_{(j \mod 2)}$. As before, let ρ denote the set of all plays. For a play $p \in \rho$, let $(p \restriction i) = \{x_{2n+(i \mod 2)}\}_{n \in \omega}$ be the projection of p to the Player i sequences(moves made by Player i). At any point in the conversation (play) p, the players obtain a finite prefix of p. Note that unlike Banach Mazur games[19], in an ME game it might be the case that $Win_0 \cap Win_1 \neq \emptyset$. Then if $p \in (Win_0 \cap Win_1)$ then both players win. Furthermore, Win_0 and Win_1 need not exhaust $(V_0 \cup V_1)^\omega$. In that case if $p \notin (Win_0 \cup Win_1)$ then neither player wins. Thus we define the following subclasses of ME games.

Definition 3 (zero-sum and non zero-sum ME game) *Let $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win_0, Win_1)$ be an ME game. Then*

- *if $Win_0 = \overline{Win_1} = Win$ (say) then $\mathcal{G}$ is said to be* zero-sum*. In that case we denote $((V_0 \cup V_1)^\omega, Win_0, Win_1)$ simply as $((V_0 \cup V_1)^\omega, Win)$,*
- *if $(Win_0 \cap Win_1) \neq 0$ then $\mathcal{G}$ is* non zero-sum*.*

A strategy s_0 for player 0 in the game, is a function from finite sequences of turns of even length to V_0^+. A strategy s_1 for player 1 is a function from

finite sequences of odd length to V_1^+. We say that a play $p \in (V_0 \cup V_1)^\omega$ conforms to a strategy s_i of player i if i plays according to s_i to generate p. A strategy s_i is winning for player i if every play p that conforms to s_i is in Win_i. Thus, no matter what the other player plays, as long as i sticks to s_i she wins. A zero-sum ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win)$ is said to be determined if either Player 0 or Player 1 always has a winning strategy.

We take strategic conversations like the examples from Section 1 to be zero-sum ME games. But the ME game framework can represent fully collaborative conversational games too, where $Win_0 = Win_1$. Examples of such fully collaborative games might be conversations involving the drafting of a jointly-authored paper. For rest of the paper, we shall concentrate on zero-sum ME games, though occasionally we will allude to ME games of other types.

An important condition on Win is whether it hinges on which of the players made a particular discourse move. We call such winning conditions decomposition sensitive. Let π denote the natural projection of $(V_0 \cup V_1)$ onto V, $\pi(v, i) = v$. Define π_ω as the extension of π into a projection of sequences in $(V_0 \cup V_1)^\omega$ onto V^ω: $\pi_\omega((v_k)_{k\in\mathbb{N}}) = (\pi(v_k))_{k\in\mathbb{N}}$ (where all the v_k belong to $V_0 \cup V_1$). π_ω is lifted to any subset A of $(V_0 \cup V_1)^\omega$ in the usual way: $\pi_\omega(A) = \{\pi_\omega(x) \mid x \in A\}$.

Definition 4 (Decomposition sensitivity) *$Win \subseteq (V_0 \cup V_1)^\omega$ is* decomposition sensitive *if $\exists W \subseteq \pi_\omega(Win)$ such that $\pi_\omega^{-1}(W) \not\subseteq Win$. Conversely Win is* decomposition invariant *if $\forall W \subseteq \pi_\omega(Win),\ \pi_\omega^{-1}(W) \subseteq Win$.*

It is easy to see that if a player i has a winning strategy in an ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win)$ she also has a winning strategy in the BM game $BM(V^\omega, \pi_\omega(Win))$: simply play the same strategy. In addition, if Win is also decomposition invariant, then the converse holds as well: if i has a winning strategy in the BM game $BM(V^\omega, \pi_\omega(Win))$ then she also has a winning strategy in the ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win)$. That's because decomposition invariance guarantees that whatever strategy she plays in $BM(V^\omega, \pi_\omega(Win))$ is bound to exist in $\mathcal{G}$. We have shown

Proposition 3 *Given an ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win)$ where Win is decomposition invariant, Player i has a winning strategy iff she has a winning strategy in the BM game $BM(V^\omega, \pi_\omega(Win))$.*

In other words, when ME games involve decomposition invariant winning conditions, they collapse to BM games, and the existence of a winning strategy is predicted by the basic theorem for BM games, which we discuss in the next section.

3.1 The Jury, constraints and meanings

But first we revisit a problem we posed for signaling games that define content in reflective equilibrium, the problem that signals have no meaning in conversations where players' interests are opposed. In our model of conversation, the opposing interests of the players do not impede communication of content but rather presuppose a set content; both players have to have a clear and defensible idea of what their opponent has committed to if they hope to win a message exchange game. While our ME games allow a player to say anything on her turn, just saying anything lacks certain important elements intrinsic to a good or winning play, and these elements end up determining this content. Players set their winning conditions vis a vis an audience that makes requirements on winning conversations. And so we need to give a more detailed model of the Jury. Henceforth, we shall develop the theory from the point of view of Player 0 alone and talk about wining conditions and strategies of Player 0 alone. This is in keeping with our concentration on win/lose zero-sum games. We note, however, that one can have the exact same arguments for Player 1 as well as the point of view of Player 1 is the dual of that of Player 0.

The Jury can either be biased towards a particular victory condition that Player 0 must guess or unbiased and accept whatever victory condition 0 chooses to play. In either case it is swayed by argument and verifies whether a particular victory condition has been met. The Jury rates each contribution by a player in individual turns or small sequences of turns with respect to whether they get a player closer to a given goal or make it more difficult to attain. We will suppose that turns are evaluated as either helping 0 achieve a particular goal, hindering her or having no effect via a function $\|\cdot\| \in \mathbb{R}$. In this paper, we will assume that the Jury is unbiased and so the Jury's and 0's conception of the winning condition coincide, though in this section we will outline some other possibilities.

The function $\|\cdot\|$ should also verify necessary conditions on good discourse like consistency and coherence, which we now describe. Discourse consistency can be defined in different ways, but it must respect the rules of valid inference. In our ME games, the rules of inference for the logical connectives and quantifiers, as well as the conventional lexicon for non logical terms, impose a notion of consistency on play. If 0, for instance, maintains in Example 1 that she proved a theorem but also that she did not prove it, she is inconsistent and the Jury will conclude she is confused. If she claims that she proved the theorem but also that if the theorem has a proof, it hasn't been found yet, she is also inconsistent. Such inconsistency precludes

her from her winning condition. We can further extend our notion of consistency by supposing further that our games *are situated* in the sense that they involve deictic reference to non-linguistic objects and properties like natural kinds. Courtroom cases typically involve extra-linguistic elements fixing the meaning of certain referring terms (imagine introducing pictures in a courtroom of a particular character, or the character himself). So consistency will involve more complex rules like how to adjust one's commitments in the light of new evidence about a natural kind or about an individual. Any violation of consistency will lead to an immediate attack by the opponent —*you just contradicted yourself so how can we (i.e. the jury) believe anything you're saying*. This is the sense of LP's closing comment in example 3. So requiring that Player 0's winning plays form a consistent set of formulas in V_0 will place constraints on the meanings of her expressions.

How does consistency affect the Jury? If Player 0 makes inconsistent contributions in the eyes of the Jury, then her contributions automatically entail that the victory conditions of Player 1 have been achieved. Given that 0 is inconsistent, her contributions entail a commitment to any content whatsoever and no information anymore; she commits to any finite sequence of discourse moves on every turn, and so 1 just needs to make some moves on each turn to achieve her winning condition.

Given that the Jury is an abstract scoring device, we may assign it powers to deduce entailments from speakers' contributions of various levels. So we consider the following form of consistency:

Definition 5 (Consistency) *A play p of an ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win)$ is* consistent *for Player 0 if the Jury does not deduce an inconsistency from $(p \restriction 0)$.*

Successful play also involves rhetorical co-operativity. Defection from the conversation is not usually a winning option for either player. The space of possibly coherent attacks on a message places constraints on the meaning of messages. An attack, or in discourse terms a Correction [1], can apply in principle to practically any word in a player's contribution, as the following adaptation of a famous example of Strawson's shows:

(4) a. 0: A man fell off a bridge.

b. 1: It wasn't a man. It was a woman, and she didn't fall; she was pushed.

c. # 1: No, John is a basket ball player.

However, not just any move can count as a coherent attack. (4c) is an incoherent discourse move and cannot be interpreted as a sensible attack or any other discourse move relating to the claim in (4a). Though player 1 can make the move in (4c), it won't do him any good *vis à vis* the Jury, and an opponent can attack it successfully as an incoherent move. Competent speakers of a language can tell quite well when an attack is coherent or not and so our players will also have to play only coherent attacks if they wish to succeed in their winning condition. Our model of the Jury below reflects this in its estimation of the type of each player; a player who is successfully attacked will suffer in the Jury's estimation of her type.

Attacks aren't the only sort of discourse move that we have to countenance. There are also rebuttals, defenses or supporting moves for claims and many others. All of these moves have coherence requirements. This is what discourse theories like that in [1] study; they make use of the lexical and compositional content of their relata to infer such relational discourse moves and check their coherence. Requiring winning plays to involve coherent discourse moves constrains the meaning a message can have.

In effect our language has a rich structure, borrowed from discourse theories like that of [1]. We have descriptions of contents of *discourse constituents* (which one can think of as elementary discourse moves) and these discourse constituents are arguments to various discourse relations like Question Answer Pair and Correction. We thus have a finite set of discourse constituent labels $DU = \{\pi, \pi_1, \pi_2, \ldots, \pi_k\}$, and a finite set of discourse relation symbols $\mathcal{R} = \{R, R_1, \ldots R_n\}$. Our vocabulary V is a finite set which consists of formulas of the form $\pi\colon \phi$, where ϕ is a description of the content of the discourse unit labeled by π and $R(\pi, \pi_1)$, which says that π_1 stands in relation R to π. Although ϕ is a formula of first order or higher order logic, it is expressed in a language that can be coded using a finite vocabulary. Using techniques like Gödel coding, we can represent any finite formula expressed with a countably infinite vocabulary of predicates, function symbols and variables within U^*, where $U = \{0, 1, 2, ..., 9\}$. Only the length of the encoding string increases (often exponentially), but it is still finite. So the assumption that V is finite is not really a restriction for linguistic analysis. Following [1], each discourse relation symbolized in V comes with constraints as to when it can be coherently used in context and when it cannot. It is these constraints that give the meanings of agents' messages and of their commitments, *irrespective of what they believe about the contents of those messages.*

With our vocabulary V now fixed, we can specify rhetorical co-operativity more precisely by noting that a sequence of conversational moves can be

represented as a graph (DU, E_1, ℓ), where DU is a set of vertices each representing a discourse unit, $E_1 \subseteq DU \times DU$ a set of edges representing links between discourse units that are labeled by $\ell : E \to \mathcal{R}$ with discourse relations. We can now define rhetorical co-operativity using the following two concepts. Given a play $p = x_0x_1x_2\ldots$ of an ME game $\mathcal{G} = (V_0 \cup V_1, Win)$ we have defined $(p \restriction 0) = x_0x_2\ldots$ to be the contributions of Player 0 in p. Let us call x_{2j} the jth turn of Player 0. For every j define the function $p_0 : \mathbb{N} \to \wp(DU)$ such that $p_0(j)$ gives the set of contributions (in terms of DUs) of Player 0 in her jth turn. That is, $p_0(j)$ is the set of DUs in x_{2j}.

Definition 6 (Coherence) *Given a play p of an ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win)$ the contribution by Player 0 is* coherent *on turn j if for all $\pi \in p_0(j)$ there exists $\pi' \in (p_0(k) \cup p_1(k-1))$ where $k \leq j$ such that there exists $e \in E_1$ such that $(e(\pi', \pi) \vee e(\pi, \pi'))$ holds.*

Definition 7 (Responsiveness) *Given a play p of an ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win)$ Player 0 is* responsive *on turn j if there exists $\pi \in p_0(j)$ such that there exists $\pi' \in (p_1(j-1))$ such that for some $e \in E_1$ we have $e(\pi', \pi)$.*

Definition 8 (Rhetorical co-operativity) *Player 0 is* rhetorically cooperative *in a play p of an ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win)$ if every turn in p by 0 is both coherent and responsive. p is rhetorically cooperative if both players are rhetorically cooperative in p.*

Note that 0 can coherently continue at turn $k+1$ her contribution from turn k, ignoring the opponent contribution in turn k. Responsiveness on the other hand, forces her to address the opponent's last turn in some way (acknowledging, correcting, answering questions, ...).

Recall our constraint NEC , a condition according to which to win a player must be able to respond to every attack. We can characterize strings that provide attacks and responses effectively in our discourse language [46]. We note that attacks come in various strengths, from corrections of fact to expositions of lies. Responses undercut such attacks.

Definition 9 (Attack) *Given a play p in an ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win)$, attack$(\pi', \pi)$ on Player 0 holds at turn j of Player 1 just in case $\pi \in p_1(j)$, $\pi' \in p_0(k)$ for some $k \leq j$, there is an $e \in E_1$ such that $e(\pi', \pi)$ and: (i) π' entails that 0 is committed to ϕ, for some ϕ, (ii) $\neg\phi$ holds.*

If the above definition holds, we shall often say that Player 1 on turn j attacks turn k of Player 0. We shall also abuse notation and denote this as $attack(k, j)$.

Definition 10 (Response) *Given a play p in an ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win)$, Response$(\pi', \pi)$ on Player 1 holds at turn j of Player 0 just in case $\exists \pi'' \in p_0(\ell), \pi' \in p_1(k)$ and $\pi \in p_0(j)$ for some $\ell \leq k \leq j$, such that Attack(π'', π') holds at turn k of 1, $\exists e \in E_1$ such that $e(\pi', \pi)$ and π implies that (i) one of 0's commitments ϕ attacked in π' is true or (ii) one of 1's commitments in π' that entails that 0 was committed to $\neg\phi$ is false.*

Again, if the above definition holds, we shall often say that Player 0 on turn j responds to Player 1's attack on turn k. We shall abuse notation and denote this as $response(k, j)$.

We can now characterize NEC as follows:

Definition 11 (NEC) *Given a play p of an ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win)$ NEC holds for Player 0 in p on turn j if $\forall \ell, k,\ \ell \leq k < j$, such that attack$(l, k)$ there exists $m, k < m \leq j$, such that response(k, m). NEC holds for Player 0 for the entire play p if it holds for her in p for infinitely many turns.*

The Jury of ME games enforces these constraints by integrating them together with the players' winning condition into what we call *the Jury winning condition*. The Jury will penalize contributions that are not coherent, and it will penalize a player that is not responsive on her turn. While being incoherent or unresponsive on a turn is not a game changer; being inconsistent is—inconsistency makes the player automatically lose. In addition, our Jury is sensitive to attacks that it deems successful; and it is sensitive to ones with no reply and thus supports NEC . The following model of the Jury with two components makes these claims concrete. Formally,

Definition 12 (Jury) *The* Jury *is a tuple $\mathcal{J} = (\{\varphi_i\}, \{P_j\}, \{c_{i,j}\})_{j \in \mathbb{N}, i \in \{0,1\}}$ where*

- *$\varphi_i \subset (V_0 \cup V_1)^\omega$ is the* Jury winning condition *for each player i where $\varphi_0 \cap \varphi_1 = \emptyset$.*
- *For each $j \in \mathbb{N}$, $P_j : \{\text{GOOD}_0, \text{BAD}_0, \text{GOOD}_1, \text{BAD}_1\} \to [0, 1]$ is a probability function with $P_j(\text{GOOD}_i) = 1 - P_j(\text{BAD}_i)$.*
- *For each $i \in \{0, 1\}, j \in \mathbb{N}$, $c_{i,j} \in (0, 1)$.*

Definition 13 (biased and unbiased Jury) *For a (zero-sum or non zero-sum) ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win_0, Win_1)$, the Jury $\mathcal{J}$ is* unbiased *if $\varphi_i = Win_i$ and it is* biased *otherwise. If for an ME game-Jury pair $(\mathcal{G}, \mathcal{J})$, the Jury is unbiased, then we say that $\mathcal{G}$ is* aligned *with $\mathcal{J}$ and it is* misaligned *otherwise.*

The Jury assigns a rating, $\|k\| \in \mathbb{R}$, to the contribution in turn k with the following constraints:

- if Player i fails to respect coherence in turn k then

$$\mathsf{COH_i(k)} = \begin{cases} -1 & \text{if } k \,\mathtt{mod}\, (i+1) = 0 \\ 1 & \text{otherwise} \end{cases}$$

- if Player i is not responsive in turn k, then

$$\mathsf{RES_i(k)} = \begin{cases} -1 & \text{if } k \,\mathtt{mod}\, (i+1) = 0 \\ 1 & \text{otherwise} \end{cases}$$

- If i is inconsistent by turn k of p, then $\mathsf{CONS_i(k')} = \mathsf{0}$ for all $k' \geq k$. Otherwise, $\mathsf{CONS_i(k')} = \mathsf{1}$

- In addition the Jury also assigns a value $\mathsf{win}_i(k)$ to every turn k of Player i as follows. Suppose p_{k-1} is the play till turn $k-1$ and suppose Player i plays u at turn k. Then $\mathsf{win}_i(k) = 1$ if $\mathcal{O}(p_k u) \cap \varphi_i \neq \emptyset$. Otherwise $\mathsf{win}_i(k) = 0$.

The Jury also maintains a probability distribution over types: $\textsc{bad}_i$ and $\textsc{good}_i$ modeling the gain or loss of credibility that i has faced so far. At each turn we write this probability as P_k, and it is defined as follows:

- $P_0(\textsc{good}_i) = 1$ and $P_{k+1}(\textsc{good}_i) = P_k(\textsc{good}_i | p_k)$, where p_k is the initial sequence of k turns in the game.

- $P_k(\textsc{bad}_i) = 1 - P_k(\textsc{good}_i)$.

- if i successfully attacks $(1-i)$ at turn k, then

$$P_k(\textsc{good}_{1-i}) = P_{k-1}(\textsc{good}_{1-i} | p_k) = c_k P_{k-1}(\textsc{good}_{1-i})$$

  where $0 \leq c_k < 1$ is a function representing the severity of punishment per single move of a player $1-i$ by the jury in the context of p_k.

These two ingredients contributes to a definition of the Jury's evaluation in the following way: $\|\tau_k\|$ of the k^{th} turn's benefits to i is given as:

$$\|\tau_k\|_i = (\mathsf{COH_i(k)} + \mathsf{RES_i(k)}) \times \mathsf{CONS_i(k)} \times \mathsf{P_k}(\textsc{good}_\mathsf{i}) \times \mathsf{win_i(k)}$$

And i's score for a play p is given as

$$\|p\|_i^{\uparrow} = \liminf_{n\to\infty} \left[\frac{1}{n} \sum_{k=1,\tau_k\in p}^{n} \|\tau_k\|_i \right]$$

Then for a sequence p, the Jury assigns a win to Player i only if $\|p\|_i^{\uparrow} > 1/2$. It is easy to see that if a necessary condition on Win_i is to win with respect to the Jury, then *Win* entails consistency. Also i must satisfy NEC , responsiveness and coherence more often than not.

A few remarks are in order at this point. First, our scoring function makes credibility in the eyes of the Jury and advancing towards the winning conditions game changers. Once a player has moved so that no continuation of her contributions are in her winning condition she has lost; similarly, once the Jury has no more confidence in the player's contributions, she cannot get a positive score on any subsequent turns and she has lost as well in the eyes of the Jury. Thus, were Janet in example (1) to admit to seeing Valentino, she has moved outside her winning winning condition of avoiding such an admission and has lost in the eyes of an unbiased Jury.

Second, our assumption $\varphi_0 \cap \varphi_1 = \emptyset$ on the winning conditions of the Jury means that the Jury assigns 'at most' one winner for the given ME game. Also, from the way the scoring function has been defined, it will always be the case that if the Jury assigns a win to Player i, it will assign a loss to Player $1 - i$. Thus even though the original ME game the players are playing might be non zero-sum, the evaluation by the Jury is always zero-sum. This is in keeping with our emphasis on zero-sum ME games in this paper.

The other situation where φ_0 and φ_1 may not be disjoint and different scoring functions should also be explored. It is an interesting question as to whether the Jury is also an integral part of a fully cooperative ME game, and what form the Jury then must take and what are the moves available to it. But we leave that question to future research. While we will concentrate here on the case of an unbiased Jury, a biased Jury, that is when $\varphi_i \cap Win_{1-i} \neq \emptyset$ for some $i \in \{0, 1\}$, might assign a win to Player i even though she loses her original ME game and conversely. All of this implies that we may have four different types of ME game-Jury pairs $(\mathcal{G}, \mathcal{J})$ depending on whether $\mathcal{G}$ is zero-sum or non zero-sum and whether the $\mathcal{J}$ is biased or unbiased. We explore these scenarios in [3].

It is easy to see the following:

Proposition 4 *If a necessary condition on Win is the Jury condition, then*

to win 0 *must respect consistency, and must satisfy* NEC *, responsiveness and coherence more often than not.*

The fact that the Jury enforces our constraints implies that the meaning of a move is largely fixed by its consistent and coherent uses in context, how it can be attacked and how it can be defended, amplified on and so on. This is a counterpart of a well-known fact about formal languages: the models of a formal language are determined by the consistency notion of the language's underlying logic or semantics. Because our players play infinite sequences and we can allow our Jury unbounded computational powers and full knowledge of a discourse move's entailments and implicatures, a player can completely specify a model for a countable first order language using Lindenbaum's procedure.[8] Suppose player i plays a move ϕ. Using the consistency and coherence requirements on winning plays, she can build a maximal consistent and coherent set of formulas from V by adding a consistent formula with a coherent relation at each turn to what she has already played. Since V contains at most ω many formulas, an infinite play of an ME game suffices to construct a maximal consistent and coherent set.[9] In fact the space $(V_0 \cup V_1)^\omega$ contains *all* maximal consistent sets for the language. Thus:

Proposition 5 *Let V be a countable first order language. Then there is a set $C \subseteq (V_0 \cup V_1)^\omega$ that consists of all plays that are consistent and rhetorically coherent and that specifies all the intended models of V.*

Not every play specifies a full model. However, the game structure itself does. And in our idealized setting, the plays that are consistent and specify models are common knowledge of the participants and of the Jury.

Our ME games thus enforce an exogenously specified notion of meaning, specified by linguistic theory. This includes implicatures, which would seem to mark an important difference with the signaling models of section 2.1. In signaling models, implicatures arise as a byproduct of co-operativity in the game's equilibrium; our model takes them to be provided by linguistic theory, and then predicts agents' attitude toward them in the conversation's continuations. We could thus think of a signaling model as one of implicature generation, while our model is one of implicature "survival".

We think the situation is more complicated for two reasons. First, on a commitment-based view such as ours, the constraints of consistency and coherence determine implicatures. For instance, Quayle's implicature K in

[8] See e.g. [9].

[9] Or a maximal consistent and saturated set.

(4) that he was comparable to John Kennedy as a politician translates in our model into the fact that it is consistent and coherent both for Bentsen to commit that Quayle committed to K and for Quayle to commit that he did not commit to K. Now as noticed earlier, a decision to exploit or to deny an implicature brings with it a commitment that the linguistic premises (co-operativity, sincerity, competence...) of the implicature's derivation hold, or do not hold. This being understood, it does not matter whether one implements those premises within a logical theory, or within a signaling game's utility profile. But linguistic constraints like coherence and consistency do further work. Consider, example (2) with a fully cooperative Bronston. Where does the implicature to the "No" answer comes from in the first place? This implicatures is fundamentally tied to coherence. Inferring "Yes" through Bronston's indirect answer makes him less coherent than inferring "No", because the "No" allows for an implicit contrastive discourse relation ("No I did not. [But] the company had one" while the yes would require an explicit marker "the company had one **too**" to infer a relation (this is moreover actually confirmed by the natural prosody of Bronston's answer). Any model would have to rely one way or another on a pragmatic theory to explain utilities and model this asymmetry.

Second, we do not think that signaling games are independent of an exogenous linguistic theory with regard to implicatures. One of the main concerns of the model in [15] is to bring conventional meaning back into signaling model, which is not innocuous by constraining the set of player types in the game. In order to capture implicatures properly, one needs to make conventional meaning a part of the signaling model. To this end [15] suggests that the set of types constitute a potential answer set to the question under discussion; hence, determining the game's context requires a pragmatic model as well.

In conclusion, both signaling games and ME game need to appeal to an exogenous theory. Signaling games give a nice implementation of the Gricean theory where linguistic considerations can often be "hidden" into the game context, whereas ME games allow a higher level form of quantification over those possible game contexts, which is crucial to account for the possibility of Gricean or non-Gricean speaker.

3.2 Some interim remarks on ME games

BM games are classically played by 2 players, as are ME games. However, conversations are not limited to two players only but may involve several players. Most of our examples in Section 4 are of that nature. Our model

can accommodate such a scenario as follows: when a conversation involves n players ($n > 2$ say), for each player i, $1 \leq i \leq n$, there is a vocabulary V_i and a winning condition Win_i. In addition, if the Win_is are such that they partition the entire space of plays $(V_1^\omega \cup V_2 \cup \ldots V_n)^\omega$ then the game is zero-sum. Intuitively each player i is playing against the 'coalition' of all the other players. Such an assumption is standard in the theory of multiplayer games.

In ME games, winning conditions are defined over sets of infinite sequences. However, our Jury witnesses actual conversations that are perforce only finite, initial prefixes of such sequences and forms a judgment concerning winners and losers of conversation on this basis. So our theory must provide a means for verifying whether a winning condition that applies to infinite strings holds or not in virtue of finite prefixes of those strings when possible. ME games whose winning conditions are not finitely verifiable will give rise to actual conversations for which a winner cannot once and for all be determined.

It may seem odd that we define a conversation's winning condition for a player using only sequences of discourse moves involving commitments. Don't agents engage in conversation typically to get their interlocutors or observers of the conversation to do some non-linguistic action? As our vocabulary can be what we can like, we can add to our vocabulary of discourse moves descriptions of non-linguistic actions or states, like *player 1 buys the goods*, which is a move that player 1 might make at the close of a negotiation. In principle we can include a description of whatever actions that are pertinent to winning conditions in a conversation.

ME games involve separate vocabularies for our two players. We have assumed that the same types of move from a vocabulary V are available to both, but our games distinguish which player plays which move by restricting 0 to play from $V_0 = \{(v, 0)|v \in V\}$ and 1 to play from $V_1 = \{(v, 1)|v \in V\}$. However, players may play with *different* vocabularies, for instance where 0 plays from the set $\{(u, 0)|u \in U\}$, 1 plays from $\{(v, 1)|v \in V\}$ and $U \subsetneq V$. That is, the moves envisioned by one player is a strict subset of the other player's set of envisioned moves. We plan to investigate this possibility in a future paper.

4 Winning conditions

Let us recap. We've argued that the analysis of strategic conversation requires a different framework with novel features, from those used in most

game theoretic analyses of conversation. We've established five necessary constraints on winning conditions in virtue of a model of the Jury—consistency, coherence, NEC from observation 1, NR from observation 2 and responsiveness; and we've shown that these constrain the meaning of the signals players use independently of the players' beliefs or preferences, in contrast to other game theoretic frameworks.

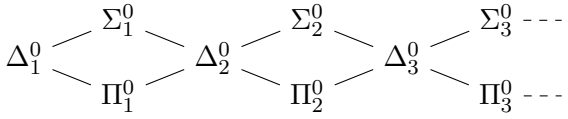
Nevertheless, while our model of the Jury evaluates individual contributions of players it does not completely determine winning conditions or the 'shapes' of conversations that depend on them. This is something we now investigate. We will characterize precisely the conversational objectives that players choose, how complex they are, whether there is a winning strategy in the game for achieving them, and, if there is a winning strategy, how complex it is. In order to do this, we need a natural way of classifying sets of infinite strings that satisfy different conversational goals. The fact that the sets of infinite strings in BM games and ME games define a topology will allow us to make use of the Borel Hierarchy, a natural measure of topological complexity, to characterize different types of winning conditions. The place of a winning condition in the Borel Hierarchy in turn determines the complexity of the winning strategy that a player should employ.[10] We will examine the structure of the winning set *Win* in a ME game when *Win* lies in the low levels of the Borel hierarchy.

We define the topology on $(V_0 \cup V_1)^\omega$ by defining the open sets to be sets of the form $A(V_0 \cup V_1)^\omega$ where $A \subseteq (V_0 \cup V_1)^*$. We shall often denote this set as $\mathcal{O}(A)$. Intuitively, $\mathcal{O}(A)$ denotes all possible ways in which a play can continue after a string $u \in A$ has been played. If A is the single string $\{u\}$ then we shall abuse notation and write $\mathcal{O}(u)$ instead of $\mathcal{O}(\{u\})$. The closed sets are as usual the complements of the open sets. Suppose for example, (2) provides an initial segment of an ME game. $\mathcal{O}(2)$ is all the ways that the conversation can continue. This is an open set over the set of plays $(V_0 \cup V_1)^\omega$ in the ME game. This topology is often called the Cantor topology on infinite strings.

Henceforth, when we say 'the space $(V_0 \cup V_1)^\omega$' or simply $(V_0 \cup V_1)^\omega$, we shall implicitly mean the topological space $(V_0 \cup V_1)^\omega$ equipped with the Cantor topology as described above. One important property of the space $(V_0 \cup V_1)^\omega$ is the following. It is a complete metric space and is second-countable; i.e., every open set in $(V_0 \cup V_1)^\omega$ can be written as a union of at most a countably infinitely many disjoint open sets [where as usual, two sets $A, B \subseteq (V_0 \cup V_1)^\omega$ are disjoint if $A \cap B = \emptyset$].

[10] See, for instance, [38] for a nice survey on infinitary games.

Using this topology, we inductively define the Borel sets, Σ^0_α and Π^0_α for $1 \leq \alpha < \omega_1$. Let Σ^0_1 be the set of all open sets. $\Pi^0_1 = \overline{\Sigma^0_1}$, the complement of the set of Σ^0_1 sets, is the set of all closed sets. Then for any $\alpha > 1$ where α is a successor ordinal, define Σ^0_α to be the countable union of all $\Pi^0_{\alpha-1}$ sets and define Π^0_α to be the complement of Σ^0_α. $\Delta^0_\alpha = \Sigma^0_\alpha \cap \Pi^0_\alpha$. Below is a schematic picture of the initial sets and their inclusion relations in the Borel Hierarchy.



Similar to the above, let us define a topology on V^ω by specifying the open sets as AV^ω where $A \subseteq V^*$. We note that since our "flattening" or projection function $\pi_\omega \colon (V_0 \cup V_1)^\omega \to V^\omega$ is an onto homomorphism, for any subset $X \subseteq V^\omega$, the Borel complexity of $\pi_\omega^{-1}(X)$ in $(V_0 \cup V_1)^\omega$ is the same as the Borel complexity of X in V^ω, where $\pi_\omega^{-1}(X) = \{s \in (V_0 \cup V_1)^\omega | \pi_\omega(s) \in X\}$. Further, π_ω maps open sets in $(V_0 \cup V_1)^\omega$ to open sets in V^ω. However, closed sets in $(V_0 \cup V_1)^\omega$ maybe mapped either to closed sets or to open sets in V^ω. By a simple inductive argument we can show that the Borel complexity of sets do not increase under the map π_ω.

Let V and Z be two vocabularies. A function $f : V^\omega \to Z^\omega$ is said to be continuous if for every open subset $B \subset Z^\omega$, $f^{-1}(B)$ is also open.

A set $A \subset V^\omega$ is said to Wadge reduce to another set $B \subset Z^\omega$, denoted $A \leq_W B$, if there exists a continuous function $f : V^\omega \to Z^\omega$ such that $f^{-1}(B) = A$.

Let V be a vocabulary. A set $A \subset V^\omega$ is said to be Σ^0_α (resp. Π^0_α) complete if $A \in \Sigma^0_\alpha$ (resp. $A \in \Pi^0_\alpha$) and for any other vocabulary Z and for any Σ^0_α (resp. Π^0_α) set $B \subset Z^\omega$, $B \leq_W A$. Intuitively, given a class of sets Γ, the complete sets of that class represent the sets which are structurally the most complex in that class.

For the Borel hierarchy, completeness can be characterized in the following simple way:

Proposition 6 ([23]) *Let $A \subset V^\omega$. Then A is Π^0_α (resp. Σ^0_α) complete if and only if $A \in \Pi^0_\alpha \setminus \Sigma^0_\alpha$ (resp. $\Sigma^0_\alpha \setminus \Pi^0_\alpha$).*

For ME games with decomposition invariant winning conditions, which are just BM games, the Banach-Mazur theorem states necessary and sufficient conditions for the existence of a strategy to achieve *Win*. The theorem

intuitively says that Player 0, the player who starts the conversation, can win if her strategy takes into account, or has an 'answer', for almost all possible situations when her turn to speak may come. That is, the set of situations that her strategy doesn't take into account must be 'small' in a sense that we define below. Example 5 from the introduction illustrates the problem. Feynmann, who confronts his students eager to stump him by beginning the conversation with "So what shall we discuss today?", automatically throws the floor open for all possible topics that might arise. He can achieve his objective only if the continuations of this initial question that do not fall within his winning condition are very few. He must have a convincing response to any possible question or topic that may be thrown at him that enables him to get to his chosen topic.

To understand this theorem, we need some definitions. A set is nowhere dense if its closure contains no non-empty open set. A set is meager if it is a countable union of nowhere dense sets. Meager sets represent sets which are 'small' in a topological sense. The complement of a meager set is a co-meager (or topologically 'large') set. A topological space is called a Baire space if the countable intersection of dense sets is dense. That is, every meager set is nowhere dense. The way we have defined our topology on V^ω and $(V_0 \cup V_1)^\omega$, both these spaces will be Baire.

Theorem 1 (Banach-Mazur [30]) *Given a BM game $BM(V^\omega, Win)$, (i) Player 1 has a winning strategy if and only if Win is a meager set; (ii) Player 0 has a winning strategy if and only if, there exists a finite string x such that $O(x) - Win$ is meager (that is, Win is co-meager in some basic open set).*

4.1 Winning conditions in conversations: Reachability and Safety

Let us start by looking at the winning condition for the conversation in Example 1, which, at least on a certain interpretation, is a very simple, decomposition invariant condition. Suppose that in order to achieve A's winning condition, someone has only to mention at some point, it doesn't matter when, that she has proved an important theorem (leaving aside the constraints NEC, NR , consistency or discourse coherence, which make winning conditions more complex). In other words, for A to win the game, the conversation must eventually contain this move. More generally, conversations in which the objective of a player is that a certain topic get mentioned

exhibit the following shape:

$$Win = Reach(R)$$

Reachability, a typical Σ_1^0 property in the Borel hierarchy, is defined as follows. Given a non-empty subset $R \subseteq (V_0 \cup V_1)$ of the elements of the vocabulary, a string x in $(V_0 \cup V_1)^\omega$ is said to reach R if the elements from R occur somewhere in x. More formally, for a string x over the vocabulary $(V_0 \cup V_1)$ we let $x(i)$ denote the ith element of x. We define $occ(x) = \{a \in (V_0 \cup V_1) \mid \exists i, x(i) = a\}$ to be the set of all the elements of $(V_0 \cup V_1)$ which occur in x. Then

$$Reach(R) = \{x \in (V_0 \cup V_1)^\omega \mid R \cap occ(x) \neq \emptyset\}$$

is the set of all strings in which at least one element of R occurs at least once. The reachability set R of example 1 is just a singleton. Since this winning condition is decomposition invariant, by the BM theorem A has a winning strategy, since the winning condition picks out an open set of strings of discourse moves.

Just as Reachability is a typical Σ_1^0 property, Safety captures some important Π_1^0 conditions and is defined as follows. Suppose S (the 'safe' set) is a subset of $(V_0 \cup V_1)$.

$$Safe(S) = \{x \in (V_0 \cup V_1)^\omega \mid occ(x) \subset S\}$$

is the set of all strings which contains elements from S alone. That is, winning plays remain in the safe set and do not move out of it. One common sort of Π_1^0 winning condition is to prevent a player from reaching a Σ_1^0 condition. Another Π_1^0 condition is consistency in a decomposition invariant ME game; such a condition can be defined without an appeal to the Jury but can exploit the consequence relation $\models$ the Jury induces: a play p is consistent iff no finite prefix p_k of p is such that $p_k \models \bot$. If this is a decomposition invariant winning condition, then this is also an example of a Π_1^0 winning condition. An example of an ME game with such a winning condition is a fully cooperative ME game, in which two agents jointly contribute to talk on a certain topic. They will want to maintain consistency in their joint view.

An alternative way of thinking about reachability and many Σ_1^0 conditions is to look at their definitions in terms of temporal logic. As our space of infinite strings is a set of linear orders, formulas of linear temporal logic (LTL) can describe some of its subsets, in particular many reachability conditions. For any element $a \in (V_0 \cup V_1)$, let the proposition p_a denote the

property of visiting or playing a. For some finite $R \subset (V_0 \cup V_1)$, the LTL defining formula for the strings that reach R is:

$$\phi_{reach(R)} = \bigvee_{a \in R} \Diamond p_a$$

where $\Diamond$ is interpreted as *eventually*.[11] A reachability formula of the form $\Diamond p_a$ is true at an index i of a sequence x ($x, i \models \Diamond p_a$) iff for some $j \geq i$, $x, j \models p_a$. A string x satisfies $\Diamond p_a$ ($x \models \Diamond p_a$) iff at the initial point 0 of x, $x, 0 \models p_a$. Safety also has an LTL defining formula for the strings that stay in S: where $\Box$ is interpreted as *always*,

$$\phi_{safe(S)} = \Box \bigvee_{a \in S} p_a$$

For a safety goal of the form $\Box\phi$, $x, i \models \Box\phi$ iff for all $j \geq i$, $x, j \models \phi$. $x \models \Box\phi$ iff $x, 0 \models \Box\phi$.

The simple Σ_1^0 winning condition of candidate A's is decomposition invariant. However, this is not so for other goals. Consider Example 2 (1) again. Justin is playing a game with a disjunction of reachability conditions as his winning condition: his goal is to get Janet either (i) to admit that she has been seeing Valentino or admit that she hasn't. These characterize an open set in an ME game where 0 is Justin. Nevertheless, unlike candidate A's, Justin's winning condition depends on Janet's making a certain commitment, and this is a decomposition sensitive winning condition. We now analyze the topological characteristics of such winning conditions in the context of zero-sum ME games, where we need only to consider one winning condition. We plan to look at the generalization of this analysis to all ME games in a subsequent paper.

We first look at a particular form of decomposition sensitive winning conditions in which *Win* depends on 0's making a contribution at each turn that depends on the history of the play. That is, 0 can never say "just anything" on her turn, if she is to win. More formally

Definition 14 (Rhetorical decomposition sensitivity) *Win is* rhetorically decomposition sensitive *iff* $\forall p \in Win$ *and for all finite prefixes* p_k *of* p *ending in a contribution by* 1, $\exists u \in V_0^*$ *such that* $\mathcal{O}(p_k u) \cap Win = \emptyset$.

Our constraints of responsiveness, coherence, consistency and NEC on winning conditions are all rhetorically decomposition sensitive conditions.

[11] For an introduction to LTL, see e.g. [24].

Proposition 7 *If Win is rhetorically decomposition sensitive, then it is meager.*

Proof Let $p \in Win$ be a winning play. Since Win is rhetorically decomposition sensitive, by definition, for every prefix p_k of p which ends with a contribution of 1 there exists a finite $u_k \in V_0^*$ such that $\mathcal{O}(p_k u_k) \cap Win = \emptyset$. Since p was arbitrary, this means the closure of Win cannot contain a dense open set. Hence, Win must be meager. □

Given the constraints enforced by the Jury, we will be mostly interested in rhetorically decomposition sensitive winning conditions in ME games. Any winning condition incorporating these constraints is a meager set.

However, unlike what the BM theorem predicts for BM games, in ME games, not all winning conditions that are meager provide a winning strategy for Player 1. 0 has a winning strategy in an ME game G just in case there is a sequence of moves p such that for every finite prefix p_k of p ending in 1's turn, there is a finite prefix u_k of plays by 0 such that $\mathcal{O}(p_k.u_k) \cap Win \neq \emptyset$. And 1 has a winning strategy in G just in case there is no such sequence p.

Consider the following abstract ME game. Suppose $V = \{a, b\}$ and suppose Player 0 loses if and only if at any point she plays b. That is, the winning set Win is

$$Win = (V_0 \cup V_1)^\omega \setminus Reach(\{(b, 0)\})$$

This is itself a rhetorically decomposition sensitive winning condition. Now, Win is a meager set. However, contrary to the BM theorem for BM games, Player 1 does not have a winning strategy in the ME game $((V_0 \cup V_1)^\omega, Win)$. That is because whatever she plays, Player 0 can always avoid playing b. In other words, the decomposition sensitivity of the ME games breaks the applicability of the Banach Mazur theorem in ME games. Player 1 cannot 'play for' Player 0 now, which she can do in the BM game. A linguistic example of such a situation is a game G where Janet from Example 2(1) is Player 0. Janet has a winning strategy in G even though her winning condition is meager.

We consider next a winning condition for 0 that depends on some finite number of contributions by 1.

Definition 15 (0-finite decomposition sensitivity) *We say that Win is* 0-finite-decomposition sensitive *if* $\exists A \subseteq V_0^*$*, such that* $\forall p \in Win\ \exists u \in A$ *such that u occurs in p.*

An instance of such a winning condition would be Justin's. Recall that Justin's objective in 2(1) is to get Janet to commit as to whether she has been seeing Valentino or not. Let Janet be Player 0 and symbolize this commitment by Janet's as $(c, 0) \in V_0$. Then Justin's winning condition is the union of open sets $\{\mathcal{O}(x.(c, 0)) \colon x \in (V_0 \cup V_1)^*\}$ and is co-meager. Thus Janet's winning condition *Win* being the complement of that of Justin is meager. Nevertheless, Janet (Player 0) has a winning strategy in such a game: never answer Justin's question directly. More generally,

Proposition 8 *If an ME game $\mathcal{G}$ has a 0-finite-decomposition sensitive winning condition, then there is no winning strategy in G for* 1*.*

Corollary 2 *There are ME games with 0-finite-decomposition sensitive winning conditions that are meager, but where* 1 *has no winning strategy.*

Similarly, we can also define 1-finite-decomposition sensitive winning conditions that depend only on finitely many moves of 1. These conditions are also meager, but Player 1 has a winning strategy, just as the BM theorem predicts.

Corollary 3 *If an ME game $\mathcal{G}$ has a 1-finite-decomposition sensitive winning condition, then there is no winning strategy in G for* 0*.*

Corollary 4 *There are ME games with 1-finite-decomposition sensitive winning conditions that are meager, and where* 1 *has a winning strategy.*

A final situation is the one of the prosecutor in Example 2(2) which has components that are decomposition sensitive, but the entire winning condition is not. The prosecutor's winning condition as described above is that Bronston must either commit to an answer or never answer P's question. Given such a winning condition, there is a winning strategy for the prosecutor: keep asking the question until Bronston commits to an answer. In fact the entire game space is the winning condition for the prosecutor. More generally:

Proposition 9 *Decomposition sensitivity of winning conditions is not preserved under union.*

Proof Let Win_1 be a decomposition sensitive winning condition and let $Win_2 = (V_0 \cup V_1)^\omega \setminus Win_1$. Clearly Win_2 is also decomposition sensitive. Indeed, because if the play is decomposed according to some play $u \in Win_1$

then Player 0 cannot win. However $Win_1 \cup Win_2 = (V_0 \cup V_1)^\omega$ is clearly decomposition invariant. □

We've now canvassed a spectrum of decomposition sensitive winning conditions in zero-sum ME games. In general, decomposition sensitivity makes ME games differ from BM games; ME games are more expressive and more complex and break the delicate link between topology and winning conditions given by the BM theorem.

Decomposition sensitivity also affects Borel complexity. If Player 0 has a winning strategy in a rhetorically decomposition sensitive ME game with a finite vocabulary of discourse moves as we have envisioned, then *Win* is complete for Π^0_2 in the Borel hierarchy. To prove this, we shall need the following lemma.

Lemma 1 ([36]) *If V is a finite vocabulary, a subset of V^ω is clopen if and only if it is of the form AV^ω where A is a finite subset of V^*.*

We have

Proposition 10 *Let $\mathcal{G} = (((V_0 \cup V_1)^\omega)^\omega, Win)$ be an ME game such that Win is rhetorically decomposition sensitive and 0 has a winning strategy. Then Win is a Π^0_2 complete set.*

Proof Call a finite play u compatible with *Win* if $\mathcal{O}(u) \cap Win \neq \emptyset$. Let a 'round' consist of a move by Player 0 followed by that of Player 1. Let U_{i0} denote the set of finite plays compatible with *Win* after 0 makes a move in round i and let U_{i1} be the set of finite plays compatible with *Win* after 1 makes a move in round i.

Claim 1 *Given the above notations, we have*

1. *$\mathcal{O}(U_{10})$ is open.*
2. *$\mathcal{O}(U_{11})$ and $\mathcal{O}(U_{i0}), \mathcal{O}(U_{i1})$ for $i \geq 2$ are open but not closed.*

Assuming Claim 1, we have

$$Win = \bigcap_{i \geq 1} (\mathcal{O}(U_{i0}) \cap \mathcal{O}(U_{i1}))$$

which is Π^0_2.

We now prove Claim 1. 1) is straightforward. For 2) first note that each of the sets U_{i0} for $i \geq 2$ and U_{i1}, for all i, is countably infinite.[12] Now, by the definition of rhetorical decomposition sensitivity, for every x in U_{i1} there exists a finite sequence $u \in V_0^*$ such that $\mathcal{O}(xu) \cap \mathit{Win} = \emptyset$. Again from the premise of the proposition which assumes that Player 0 has a winning strategy in Win, it must be the case that there is a finite sequence $v \in V_0^*$ such that $\mathcal{O}(xv) \cap \mathit{Win} \neq \emptyset$. This means U_{i1} has a countably infinite subset $\hat{U}_{i1}$ such that for every $u, v \in \hat{U}_{i1}$, neither u is a prefix of v nor v is a prefix of u. Thus $\mathcal{O}(U_{i0})$ cannot be written in the form $A((V_0 \cup V_1)^\omega)$ such that A is a finite subset of $(V_0 \cup V_1)^*$. Hence, by Lemma 1 $\mathcal{O}(U_{i1})$ cannot be clopen which implies $(\mathcal{O}(U_{i0}) \cap \mathcal{O}(U_{i1}))$ is not clopen either. However $(\mathcal{O}(U_{i0}) \cap \mathcal{O}(U_{i1}))$ is open by definition. This proves the claim.

Next, this intersection defining Win is not itself open. To see this note that the open sets U_{ij} form chains of subsets ordered by $\subseteq$. Since $(V_0 \cup V_1)^\omega$ is a complete metric space, it contains the limit points of the intersections of each of these chains. Thus, Win, being non-empty, itself contains at least one limit point and is hence not open.

Next, we show that Win is not closed either. Let Win_k denote the kth stage in the above construction of Win. Formally, let

$$\mathit{Win}_k = \bigcap_{1 \leq i \leq k} (\mathcal{O}(U_{i0}) \cap \mathcal{O}(U_{i1}))$$

Thus, $\mathit{Win} = \bigcap_{k \geq 1} \mathit{Win}_k$. By the argument above Win_k contains countably many disjoint open subsets. Also, from the above argument using the fact that Win is rhetorically decomposition sensitive and that 0 has a winning strategy, each disjoint open subset A of Win_k, in turn, has countably many disjoint subsets in Win_{k+1}. Continuing this argument, we see that Win contains an uncountably many disjoint subsets. Now suppose, Win is closed. Then each of the uncountable disjoint subsets is closed. Thus $((V_0 \cup V_1)^\omega \setminus \mathit{Win})$ must contain an uncountable set of disjoint open sets. But that is a contradiction, since $(V_0 \cup V_1)^\omega$ is second-countable, every open set is a disjoint union of at most countably many open sets.

Finally, we show that Win is not Σ_2^0. For that we show that $((V_0 \cup V_1)^\omega \setminus \mathit{Win})$ cannot be Π_2^0. Suppose, $((V_0 \cup V_1)^\omega \setminus \mathit{Win})$ were indeed Π_2^0. Then, since $((V_0 \cup V_1)^\omega \setminus \mathit{Win})$ is not Σ_1^0, by definition, it has to be a countable intersection of Σ_1^0 sets. That is, $((V_0 \cup V_1)^\omega \setminus \mathit{Win}) = \bigcap_{k \geq 1} A_k$ where each A_k is a Σ_1^0 set. Since this intersection is non-empty—i.e., $((V_0 \cup V_1)^\omega \setminus \mathit{Win})$ is

[12] This holds assuming 0 has a countable number of moves in her first turn. Otherwise, only U_{10} is finite and the arguments continue to hold.

non-empty, $\{A_k\}_{k\geq 1}$ can be partitioned into chains of subsets, each of which is ordered by $\subset$. Using, once again the fact that $(V_0 \cup V_1)^\omega$ is a complete metric space, we must have that the limit points of each of these chains are in $(V_0 \cup V_1)^\omega$. Also, these points are disconnected. Now, since $((V_0 \cup V_1)^\omega \setminus Win)$ is non-empty, it is the union of a subset of these limit points. However, since these points are disconnected, it means $((V_0 \cup V_1)^\omega \setminus Win)$ cannot contain an open set. This is a contradiction, since $((V_0 \cup V_1)^\omega \setminus Win)$ contains, for example, $\mathcal{O}(xu)$, where v is as defined above. Hence, $((V_0 \cup V_1)^\omega \setminus Win)$ cannot be Σ_2^0.

Thus, we have shown that $Win \in \Pi_2^0$ and Win is neither open nor closed nor a Σ_2^0 set. Hence, using Proposition 6, we have that Win is complete for Π_2^0. □

All rhetorically decomposition sensitive winning conditions, including consistency, responsiveness, coherence and NEC, are Π_2^0, if Player 0 has winning strategies. But consistency often (and always if it's the only constraint) has a winning strategy (just repeat one consistent proposition whenever it's your turn) and similarly responsiveness and coherence also have winning strategies. So as a corollary we have:

Corollary 5 *Consistency, responsiveness, coherence as constraints on Win are all Π_2^0 if Player 0 has a winning strategy.*

4.2 Winning Conditions: co-Büchi

The next class of sets in the Borel Hierarchy that are interesting to us are the Σ_2^0 sets. The so called co-Büchi sets form typical sets of this class and are defined as follows. Suppose C is a subset of $(V_0 \cup V_1)$ (the 'co-Büchi' set). Then

$$co\text{-}Büchi(C) = \{x \in (V_0 \cup V_1)^\omega \mid \inf(x) \subseteq C\}$$

where $inf(x) = \{a \in (V_0 \cup V_1) \mid \forall i, \exists j > i, x(j) = a\}$ is defined to be the set of all the elements of $(V_0 \cup V_1)$ which occur infinitely often in x.

In terms of LTL formulae, the co-Büchi condition may be viewed as follows. Let $C \subseteq (V_0 \cup V_1)$ be the co-Büchi set. Then

$$\phi_{co-Buchi(C)} = \Diamond\Box \bigvee_{a \in C} p_a$$

Classic examples of co-Büchi conditions are those with strings that eventually contain only elements of C or eventually settle down in C. That is,

the strings eventually get stuck in the safe set C. Example 5 is a motivating example for a conversation with a Σ_2^0 winning condition: Feynmann had to respond to his students' questions and in a coherent way lead them eventually to the topic that he wanted to discuss.

The winning condition that a conversation be finite is also a co-Büchi condition. A finite conversation is easily modeled in the ME framework; the initial segment in which the agreement is reached is then succeeded by an infinite sequence of "null" moves that keep the content of the last move. Indeed, in the non zero-sum setting, that is, where the interests of the players are not strictly opposed, there is a close connection between agreement winning conditions and finiteness. If the only goal of the exchange is to achieve a fixed point in which the dialogue stays within this information state forever after, the conversation should stop once the terms of the exchange and the agreement are common knowledge. Being rational agents, our players will stop once they acquire the mutual knowledge that that state has been achieved and that nothing will take them out of it.[13]

Bargaining agreements or agreements on some permanent exchange of goods, which could also be information are naturally Σ_2^0 conversations even in the absence of these constraints— For e.g. any information seeking conversation in which 0 has the goal of acquiring agreement about some intellectual issue ϕ, like a Socratic dialogue also has the structure of a Σ_2^0 winning condition. [47] calls these inquiry dialogues. Co-Büchi conditions distinguish between provisional and real agreement. In a provisional agreement, an agent provisionally may acknowledge another's contribution and agree to a bargain but later take the acknowledgment and the agreement back. If *Win* of the conversational game consists only in reaching a provisional agreement, *Win* is clearly Σ_1^0, as it does not constrain what happens after the provisional agreement is reached. Real agreement is different. Once attained between two agents, the agents do not deviate from it in any further conversation; no conversational moves take them out of that state of agreement, as required for a co-Büchi condition.

Such a "lasting agreement" co-Büchi winning condition for 0 is 0-decomposition sensitive and 1-decomposition sensitive. In the zero-sum setting, in the absence of any constraints, Corollary 3 tells us that a lasting agreement winning condition has no winning strategy for 0; 1 always has a non empty move of disagreeing for any possible continuation by 0. So, no conversational goal of

[13] In principle, participants could continue acknowledging each other's acknowledgments *ad infinitum*. But such acknowledgments wouldn't serve any purpose. For a discussion see [45].

extracting a binding oath from an opponent can succeed, unless additional constraints are imposed. And similarly, a winning condition for 0 that the conversation be finite has no winning strategy—1 can always prolong the game by talking when it is her turn.

It is for this reason that "lasting agreement" winning conditions cannot happen without relaxing the assumption that players are playing a zero-sum game. There is an exception, however. The Jury can be an "arbitrator", imposing agreement when the opponent 1 no longer has any counter arguments to rebut 0's arguments for a particular position or exchange; the lack of counter arguments makes 1's objections not credible, thus lowering 1's score eventually leading to 0's winning condition. Thus, with the Jury's constraints, 0 wins a Σ_2^0 goal iff 1 has eventually no more arguments against a certain proposition ϕ, where ϕ may describe a bargain or topic of discussion.

Co-Büchi conditions also characterize goals in which 0 repeatedly attacks 1 eventually to reduce the opponent's score in the eyes of the Jury [46]. This Σ_2^0 goal is not 1-decomposition sensitive. (3) is such an example. Let LD be Player 0 in an ME game. In the *voire-dire* transcript, 0 repeatedly returns to the question as to whether the defendant Tzeng was responsible for severing a nerve in a patient's hand; he seemed prepared to revisit the theme indefinitely until he exposes that the expert witness D, or 1 in this game, was covering up for a fault of the defendant. Repeatedly questioned, 1 replies each time in the play up to (3c,d) that Tzeng was not at fault. In the Jury's eyes, 0's questioning had little effect; the Jury's probabilities assigned to the types of 0 and 1 did not shift, and 0 was no closer to his winning condition in getting the court to agree with him that 1 was not an impartial witness. However, at (3c,d), 1 contradicts his previous testimony by agreeing to 0's loaded question, and his attempts to backtrack and correct his mistake are successfully attacked by 0 in (3h). At this point, 0 has achieved his goal. We note that our model of the Jury needs refinement in that it does not take account of successful retractions in the face of inconsistency, and so we cannot really predict the counterattack at (3h). We plan to address this in future work.

4.3 Büchi winning conditions

The complementary condition of a Co-Büchi condition, the Büchi condition, is the equivalent of (infinite) iterated reachability, and is a condition that is not expressible on finite strings. Büchi conditions are Π_2^0, which we've already met. Suppose B is a subset of $(V_0 \cup V_1)$ (the 'Büchi' set). Then

$$\mathit{Büchi}(B) = \{x \in (V_0 \cup V_1)^\omega \mid \inf(x) \cap B \neq \emptyset\}$$

is the set of all strings which contain infinitely many elements of B or equivalently which visit B infinitely often. A Büchi set is Π^0_2 in the Borel hierarchy. In conversations where Player 0 has a Büchi winning condition, she will win if she always has a path to B and revisits B infinitely often. 0 can play for a Büchi condition and allow 1 a reachability or Σ^0_1 condition on his play. In such a game, Player 0 can continue to return to her chosen and preferred states infinitely often, reiterating a point or set of points that she wants to make (once again, a finite conjunction of Büchi conditions is also Büchi). All rhetorically decomposition sensitive winning conditions are Büchi conditions as we've already shown.

Büchi conditions can be expressed using LTL formulas. Let $B \subseteq (V_0 \cup V_1)$ be the Büchi set of states. Then

$$\phi_{Büchi(B)} = \Box\Diamond \bigvee_{a \in B} p_a$$

If 0's winning condition means revisiting a set of states B infinitely often, it must be for some purpose other than agreements on exchanges of goods or information, for once lasting agreement is achieved, there is no point in revisiting that agreement. On the other hand, a Büchi condition can be effective in debate. Political debates like those evoked in example (6) exemplify a Büchi condition. Such a condition is more difficult to achieve if our rhetorical constraints are imposed on acceptable discourse sequences, because it means that any play by the opponent must still enable the player to have a rhetorically cooperative path to return to B. But a practiced debater can have such a strategy.

Let's now take a closer look at the analysis of one of our examples involving a Büchi winning condition, (4). Our excerpted example was a turning point in the Vice-Presidential debate. Quayle was recurrently questioned about his inexperience and one of Quayle's goals in the debate was to continue to fend off these attacks with the defense that despite his youth he had the talent and experience of a good Vice-Presidential and Presidential candidate. In effect this is a Π^0_2 winning condition and an instance of NEC. We concentrate on this goal here and assign Quayle to be Player 0. Up to the exchange in 4, we can assume that Quayle had not made any disastrous moves, had remained consistent, responsive and replied to attacks and that the Jury's assignment to GOOD_0 had not suffered that much. His play had produced an initial segment of strings in *Win*. That is, we assume that the play p up until (4) is such that $\|p\|$ was above 0, though not significantly above.

It would seem 0 had a clear winning strategy. What went wrong? To describe the exchange in (4) in detail, we need as basic vocabulary for both V_0 and V_1 which we describe below. Given a play p in $(V_0 \cup V_1)^\omega$ we have

- An attack move, $\mathit{attack}(\pi', \pi)$, where π' and π are contributions of ith and jth turns of players 0 and 1 respectively for $i \leq j$, meaning that move π attacks move π'.
- Descriptions of the content of basic moves $\pi\colon \phi$ and $\pi'\colon \psi$,
- a commentary move, $\mathit{comment}(\pi', \pi)$ where a player expresses an opinion in π about move π',
- and a question answering move QAP.

(4a) is a QAP move to a question about his Presidential qualifications. But the content of (4a) is ambiguous. Quayle might have just intended (4a)'s literal meaning—that he was equal in governmental experience to John Kennedy as a candidate for President. But he might also have intended, and probably did intend by mentioning a famous President, to have, in light of his winning condition, the audience and Jury draw a direct and positive comparison between himself and Kennedy with regards to the kind of President he might become. In response Bentsen plays $\mathit{attack}(a, b)$ with $b\colon \phi$, where ϕ contradicts the implicated direct comparison. This is exactly what he should have done; to try to get the Jury to lower their estimation of Quayle's GOOD type. At this point Quayle should have counter-attacked with another *attack* move, as NEC requires. But instead, Quayle plays a weak $\mathit{comment}(b, c)$ with $c\colon \psi$ in the subsequent turn; and then Bentsen plays another successful $\mathit{attack}(c, d)$ with $d\colon \chi$, where χ says that it was Quayle brought up the comparison and thus opened himself up to attack—hence, $\mathit{attack}(a, b)$ was perfectly fair. Even after that Quayle still did not attack Bentsen on that point.

The Jury penalized Quayle severely for his repeated failure to assume and defend the consequences of an implicature he most likely intended given his winning condition, setting $c_{\mathit{comment}(b,c)} = 0$. In fact, the semantics of $\mathit{comment}(b, c)$ implies that Quayle accepted Bentsen's attack and its implicatures, which is that he was trying to insinuate that he was a potential another Kennedy and that he knew that he wasn't. In effect the Jury punishes Quayle for a bit of dishonesty and hence $P_d(\text{GOOD}_0) = 0$. Thus, at this point, $\|p.abcd\| = 0$, and Quayle could do nothing in the remainder of the debate to get $P_i(\text{GOOD}_0) > 0$. Given these assumptions, our model predicts that Quayle lost the debate with this one move, though we should ideally

refine the notion of the Jury's probability distribution to allow Quayle to have a chance at recovery.

4.4 Muller winning conditions

The final type of winning conditions we study are called Muller conditions. A Muller condition is defined as follows. Suppose we are given a set $\mathcal{F} = \{F_1, F_2, \ldots, F_n\}$ of subsets of $(V_0 \cup V_1)$ (the Muller sets). Then $Muller(\mathcal{F}) = \{x \in (V_0 \cup V_1)^\omega \mid inf(x) \in \mathcal{F}\}$ is the set of all strings which eventually (after a finite point) get stuck in one of the Muller sets in $\mathcal{F}$.

A Muller winning condition is a boolean combination of Büchi and Co-Büchi conditions. In terms of temporal logic formulae this can be seen as follows. Let $\mathcal{F} = \{F_1, F_2, \ldots, F_n\}$ be the set of Muller sets where each F_i is a subset of $(V_0 \cup V_1)$. Then

$$\phi_{Muller(\mathcal{F})} = (\phi_{co\text{-}B\ddot{u}chi(F_1)} \vee \ldots \vee \phi_{co\text{-}B\ddot{u}chi(F_n)}) \wedge (\phi_{co\text{-}B\ddot{u}chi(F_1)} \Rightarrow$$

$$\bigwedge_{a \in F_1} \phi_{B\ddot{u}chi(\{a\})}) \wedge \ldots \wedge (\phi_{co\text{-}B\ddot{u}chi(F_n)} \Rightarrow \bigwedge_{a \in F_n} \phi_{B\ddot{u}chi(\{a\})})$$

Since Muller conditions extend Büchi conditions, Muller conditions are not compatible with the goal of exchanging goods. Nevertheless, there are real life conversations with Muller "winning" conditions with multiple states, in which the participants revisit the states indefinitely often. In fact conversations with Muller winning conditions are commonplace. For instance, examples (3) has both Π^0_2 and also Σ^0_2 components to his winning conditions, as the Jury requires that they obey rhetorical co-operativity, NEC and consistency. Once a Π^0_2 winning condition is combined with a Σ^0_2 requirement, the result is a Muller winning condition.

Proposition 11 *If* 0 *must obey a rhetorically decomposition sensitive condition with a* Σ^0_2 *objective, then her winning condition is Muller.*

There are also examples of conversations with Muller winning conditions, without considering the Jury enforced constraints of consistency, coherence, responsiveness and the like. For instance, a conversation between two partners who have lived for a long time together and who are quite old. After a certain point 0 always attempts to go through the same conversational moves, so that they can revisit the same memories, and touch on the same themes, laugh at the same jokes. 0 asks the same questions to get the same answers. To quote John Prine from the song *Far from Me*, *a question ain't*

really a question if you know the answer too. In the song 1 plays along for a while, though she "waits a little too long," to laugh at the same, repeated jokes. In the end 1's goal is to break the cycle of repeated conversational moves by 0. Assuming that our conversational agents are rational, the goal of such a conversation is not information exchange or some sort of persuasion; it is something else like venting one's emotions albeit indirectly, reliving an experience, or conveying some other non literal message.

4.5 Relation to Gale-Stewart games

Another type of infinite games that have been extensively studied in the literature of Logic and Computer Science is called a Gale-Stewart game (GS game). A GS game is similar to a BM game in that the play of such a game is again an infinite sequence x over a vocabulary V. However, whereas in a BM game the players take turns in playing finite non-empty *sequences* of letters from V, in a GS game, the players can play only single letters (from V) in each turn. In other words, the turn-structure of the play is built inherently into the dynamics of the game. Thus a GS game over a vocabulary V is a tuple $G = (V, Win)$ where Win, as before, is a subset of V^ω. The notions of a strategy, winning strategy, determinacy etc are defined as a BM game. Every GS game (V, Win) where Win is Borel is determined [29]. However, the rigid turn structure of GS games precludes a characterization of the winner in terms of the topological properties of the winning set unlike in a BM game (thanks to the BM theorem). GS games in fact capture our ME games with decomposition sensitive winning conditions, while the BM games are isomorphic to the decomposition invariant ME games.

The ME conversational games we proposed in this paper were developed by imposing a turn structure on BM games. Alternatively, we could have developed them as GS games (as the turn structure is built in) on the vocabulary $V^+ = V^* \setminus \{\epsilon\}$. Thus, the players take turns in playing elements from V^+ (which are finite non-empty sequences in V) forming an infinite sequence in V^ω. A play can thus be viewed as a sequence in $((V_0^+ \cup V_1^+)^\omega$ where $V_i = V \times \{i\},\ i \in \{0, 1\}$ as before. Win once again, is a subset of $((V_0^+ \cup V_1^+)^\omega$.

Exploiting the correspondence between GS games and decomposition sensitive ME games allows us to apply the Büchi-Landweber theorem, to infer the memory requirements of a winning strategy. To state this theorem we need a little background on what exactly is the memory of a strategy.

A strategy s_0 for Player 0, in a GS game (V, Win) is a function $s_0 : (V_0^+V_1^+)^* \to V^+$ and a strategy s_1 for Player 1 is a function $s_1 : (V_0^+V_1^+)^*V_0^+ \to$

V^+. A strategy s_i of player $i \in \{0,1\}$ is said to be finite memory if there exists a finite automaton with output, $\mathcal{M}_i$ which dictates s_i. More formally let, $\mathcal{M}_i = (M_i, m_i^0, \delta_i, O_i)$ where M_i is a finite set of states (the memory of s_i), $m_i^0 \in M_i$ is the initial memory, $\delta_i : V \times M \to M$ is the transition function and $O_i : V \times M \to V^+$ is the output function. Define the extended transition relation $\widehat{\delta_i} : V^+ \times M \to M$ from δ_i inductively as usual. That is, $\widehat{\delta_i}(v, m) = \delta_i(v, m)$ and $\widehat{\delta_i}(xv, m) = \delta_i(v, \widehat{\delta_i}(x, m))$. Then for every finite play xv that ends in Player $(1-i)$s turn, $s_i(xv) = O_i(v, \widehat{\delta_i}(xv, m_i^0))$. s_i is memoryless or positional if M_i is a singleton. A positional strategy can be represented as a function $s_i : V \to V^+$. Now, the Büchi-Landweber theorem can be stated as follows

Theorem 2 ([8]) *Let $G = (V, Win)$ be a Gale-Stewart game such that V is finite. If the Borel complexity of Win is at most a boolean combination of Σ_2^0 and Π_2^0 sets (that is, Win is at most Muller), then one of the players always has a finite memory winning strategy.*

Coming back to conversations, this means that to win an ME game with a decomposition sensitive winning condition having a low Borel complexity, a player may require memory. However, if we restrict the vocabulary of the players by fixing a bound on the length of the sentences they can utter in each round and if the winning condition is not 'too complicated', a finite amount of memory suffices. Here is an example illustrated with a particularly simple ME game with $V = \{a, b\}$:

(5) Consider an ME game $\mathcal{G} = (V, Win)$ where 0 achieves *Win* iff she plays a and b infinitely often. That is, she does not play either all a's or all b's forever after a certain point. 1's moves do not matter.

0 has a winning strategy for $\mathcal{G}$: alternate between a and b in successive moves. But to play this strategy, she has to remember what she did on her prior turn to achieve it. So she would need to have at least one cell of memory for a winning strategy. As another more linguistically sophisticated example, any winning condition that incorporates NR would require (2 cells of) turn memory.

As finite automata are one of the most tractable algorithmic objects, this suggests to us an ambitious but exciting future project: Given a debating situation between two (or more candidates) where the goals of the candidates can be represented as low-complexity Borel sets, predict the winner and design a winning strategy for her.

4.6 Beyond first order definable winning conditions

There are ME games with intuitive winning conditions more complex than Muller conditions. The constraint CNEC is a modification of NEC, and it is quite intuitive. It is a winning condition that various instances of the Jury might employ to judge say a political or philosophical debate. The motivation behind CNEC is that there are many times where a player cannot adequately reply to an attack, and yet she can still retain intuitively a strategy for achieving her objectives by counter-attacking more often. In the literature on argumentation, it is often assumed that any attack on an attack renders it moot and ineffective [12]. CNEC requires that a string is winning for 0 if intuitively 0 defends her commitments more successfully than 1 defends hers; in other words 0 has more successful unrefuted attacks on 1 than vice versa. Note that this condition is beyond the expressive capacity of first order logic over linear orders; we need at least first order logic with counting quantifiers [26]. But in fact, we need slightly more than this for the following reason: such a winning condition is impossible for 0 to attain at every stage in the game; on her turn 1 can always pile on the attacks that have yet to be answered by 0. This is a *tail condition* in the sense of [10].

Definition 16 (CNEC) CNEC *holds for Player* $i \in \{0,1\}$ *on turn* j *of a play* p *if there are fewer attacks on* i *with no response in* p_j *than for* $1-i$. CNEC *holds for Player* $i \in \{0,1\}$ *over a play* p *if in the limit there are more turns of* p *where* CNEC *holds for* i *than there are turns of* p *where* CNEC *holds for* $1-i$.

We note that the Jury as we have designed it verifies CNEC . The set CNEC $\subset (V_0 \cup V_1)^\omega$ is defined as

$$\mathsf{CNEC} = \left\{ p \in (V_0 \cup V_1)^\omega \;\mid\; \liminf_{n\to\infty} \frac{\textit{good attacks by 0 in } p_n}{\textit{good attacks by 1 in } p_n} \geq 1 \right\}$$

That is, Player 0 has the upper hand over her opponent more often. Now by the definition of lim inf we have the following sequence of equalities

$$\begin{aligned}
\mathsf{CNEC} &= \bigcap_{N>0} \left\{ p \in (V_0 \cup V_1)^\omega \;\mid\; \exists m > 0,\ \forall n > m\ \frac{\textit{good attacks by 0 in } p_n}{\textit{good attacks by 1 in } p_n} > 1 - 1/N \right\} \\
&= \bigcap_{N>0} \bigcup_{m>0} \bigcap_{n\geq m} \left\{ p \in (V_0 \cup V_1)^\omega \;\mid\; \frac{\textit{good attacks by 0 in } p_n}{\textit{good attacks by 1 in } p_n} > 1 - 1/N \right\}
\end{aligned}$$

Now note that the set

$$\bigcap_{n \geq m} \left\{ p \in (V_0 \cup V_1)^\omega \;\mid\; \frac{\textit{good attacks by 0 in } p_n}{\textit{good attacks by 1 in } p_n} > 1 - 1/N \right\}$$

is closed. So, CNEC is a Π^0_3 set. Using ideas from [10], since CNEC is a tail condition, we can also show that CNEC cannot be expressed as a set of Borel complexity ≤ 2. That is, CNEC is Π^0_3 hard. We thus have

Proposition 12 CNEC *is a* Π^0_3 *complete set*

Conditions that are at least Σ^0_3 or Π^0_3 in Borel complexity do not have finitely satisfiable conditions; nor are they expressible using LTL formulae or even first order formulas of the language of linear orders [31]. Thus the full constraint CNEC is not expressible as a first order formula over linear orders.

4.7 ME winning conditions on finite strings revisited

The examples of winning conditions that we have examined are all expressible as formulas of linear temporal logic (LTL), whose semantics uses infinite, linearly ordered sequences of evaluation points for formulas. We've argued that players should choose winning conditions for which they have a winning strategy, allowing them to stay in the set of plays picked out by *Win* no matter how long the conversation goes on.

But all actual conversations are resolutely finite. So what can our characterization of winning conditions say about *actual* conversations? To address this issue we need the following definition of finite satisfiability.

Definition 17 (Finite satisfiability and refutability) *Given an ME game* $\mathcal{G} = ((V_0 \cup V_1)^\omega, \mathit{Win})$ *a sequence* $x \in (V_0 \cup V_1)^*$ finitely satisfies *Win if* $\mathcal{O}(x) \subseteq \mathit{Win}$ *and it* finitely refutes *Win if* $\mathcal{O}(x) \cap \mathit{Win} = \emptyset$. *Win is said to be* finitely satisfiable *(resp.* finitely refutable*) if there there exists such a sequence x that finitely satisfies (resp. finitely refutes) Win.*

Thus if *Win* is finitely satisfiable, there is a finite sequence x which already 'attests' to the satisfiability of *Win* and if x has been played so far, Player 0 can already be declared as the winner. It is easy to see that our LTL condition for reachability $\Diamond\phi$ is finitely satisfiable whereas its negation, $\Box\neg\phi$ is finitely refutable. For $\Diamond\phi$, Player 0 can, in any finite turn i 'play' a basic move with a description ϕ so that no matter how the play continues after that, she is assured a win. If ϕ is a Σ^0_1 condition and it is true in σ,

then it is true in every extension of σ; conversely, if ϕ is a Π^0_1 condition and it is false in σ, then it is false in every extension of σ. This of course is a direct consequence of the quantificational structure of the descriptions of these sets and the property that existential formulae are preserved in all extensions of a finite linear order. Finite satisfiability may provide the Jury with a criterion to evaluate an ME game. For example, if for an ME game *Win* is finitely satisfiable, and at some point, the finite play x finitely satisfies *Win*, then the Jury can already assign the win to Player 0 (and similarly for finite refutability). The stable verification or refutation of ϕ in a finite string also has a topological characterization. Consider all those infinite strings s such that all of the finite prefixes of s satisfy ϕ. The set of all such strings S is a model for ϕ; if S is open, then any game with ϕ as a winning condition is verified at a finite point, the point where S starts to branch. If all continuations of a finite prefix s verify or fail to verify ϕ, then the game is already decided at s.

Thus, in Example 2, the prosecutor has achieved his Σ^0_1 goal of extracting at least a defeasible commitment from Bronston that he had no account. Under the absence of further constraints, he has achieved his goal and 'won'. However, the game can and does go on in coherent, responsive and consistent plays, obeying NR and NEC. Such plays might refute the initial simple Σ^0_1 objective of the prosecutor. Thus, given that the prosecutor's choice of winning condition was one the opinionated Jury found persuasive, it and the unbiased Jury would award the prosecutor a victory. However had the opinionated Jury required a stronger winning condition, on which the prosecutor had to extract a non-defeasible commitment from Bronston, it would have assigned Bronston the win.

Next, in Example 3, LP's goal cornering leading D into an inconsistency yields LP a win, given our specification of the Jury. It is easy to see that consistency is finitely refutable in the above sense, or dually, inconsistency is finitely satisfiable. Hence the Jury had all the evidence it required to grant the win to LP.

How about Büchi, co-Büchi and Muller winning conditions? Note that a Büchi condition is of the LTL form $\Box\Diamond\phi$ and is 'truly infinitary' in that, it is neither finitely satisfiable nor refutable in the above sense. A similar observation holds for co-Büchi and hence Muller winning conditions as well. Since, Büchi and co-Büchi are in the 2nd level of the Borel hierarchy, we can conclude that finite satisfiability does not hold for sets in the 2nd level of the Borel hierarchy.

Σ^0_2 or Π^0_2 goals can be true on a finite sequence of evaluation points, but they are unstable. A Σ^0_2 goal like an agreement may be reached in one finite

sequence but then falsified in a finite continuation of that sequence, only to be reinstated in another continuation. This instability is reflected in the failure in our finitary semantics of certain LTL entailments, which shows us that these conditions are not really captured on finite strings. For instance a formula of the form $\Box\Diamond\phi$ can be satisfied in a finite sequence x of length 1 (such that ϕ is true at the first index) but its LTL consequence $\bigcirc\Box\Diamond\phi$, where $\bigcirc$ is the next-time operator, is clearly false. The unstable behavior of conditions at the second level of the Borel Hierarchy also reflects some of our linguistic observations—for instance, the difficulty of establishing lasting agreements over a finite fragment of conversation. Another example of this instability is this: NEC, the necessary condition on winning, may be falsified "unjustly", if 0 does not get the chance to reply to an attack. Perhaps for this reason, if the opponent launches an attack, most observers would require as fair that the attacked agent have the right of reply. Indeed all of our rhetorical constraints are unstable as can be seen from Proposition 11.

Even Σ^0_1 conditions of the form $\Diamond\phi$ could fail to be true over a particular finite sequence but true in some continuations. But for goals at the second level of the Borel Hierarchy and higher, the goals cannot be determined one way or the other by any finite prefix of an infinite play. So to capture the link between winning conditions at the second level of the Borel Hierarchy and beyond, we propose a stronger notion for the finite analysis of such sets which we call 'strategic finite satisfiability'.

Definition 18 (Strategic finite satisfiability and refutability) *Given an ME game $\mathcal{G} = ((V_0 \cup V_1)^\omega, Win)$ a sequence $x \in (V_0 \cup V_1)^*$* strategically finitely satisfies *Win if $\mathcal{O}(x) \cap Win \neq \emptyset$ and Player 0 has a strategy s_0 such that any play $p = xu \in (V_0 \cup V_1)^\omega$ conforming to s_0 is in Win. Similarly x* strategically finitely refutes *Win if $\mathcal{O}(x) \subsetneq Win$ and Player 1 has a strategy s_1 such that any play $p = xu \in (V_0 \cup V_1)^\omega$ conforming to s_1 is not in Win. Win is said to be* strategically finitely satisfiable *(resp.* strategically finitely refutable*) if there there exists such a sequence x that strategically finitely satisfies (resp.* strategically finitely refutes*) Win.*

Strategic finite satisfiability gives the Jury a handle to 'judge' plays for winning conditions which belong to or are beyond the 2nd level of the Borel hierarchy. For example, if in a play Player 0 has been successfully able to respond to every attack so far (and has also been consistent), the Jury might be convinced that she has a winning strategy for CNEC and grant her a victory, given a finite conversation. Thus, a finite sequence may give a Jury evidence that a Σ^0_2, Π^0_2 winning condition or higher will be stably

verified or refuted on the continuations of the observed, finite sequence that the strategies of the players make possible. A finite conversation can give evidence that a player has a winning strategy in an infinitary game. Clinton, for instance, repeatedly demonstrated his ability to return to economic themes despite all sorts of attempts and events during the course of the Presidential campaign that could have derailed him. On the other hand, a player with a higher goal can stumble so badly that a finite conversation seals his or her doom. This happened in the debate from which example (4) is excerpted; after Bentsen's attack, our model predicts that Quayle's objectives were foiled, no matter how he continued.

5 Conclusions and prospects

We have proposed a model for conversations in a strategic setting, ME games, building on BM games. ME games are infinitary games, and we have argued that infinitary games are essential to the analysis of strategic conversation. The crucial difference between strategic and fully cooperative contexts is that in strategic contexts, but usually not in cooperative ones, the players must behave as if the game is infinite, because they do not know who is going to have the "last word". In Gricean contexts, both players are trying to achieve a joint goal, and all things being equal, they want to achieve it efficiently. That efficiency is what limits the length of the game. Signaling games are very good at capturing these limited (especially one-off) interactions. But as soon as strategic considerations are introduced, this efficiency goes out the window.

ME games show, in distinction to BM games, the importance of turn structure in dialogue; the goals in a strategic conversation typically rely on particular contributions from each participant. In conversation, it often really matters who says what. And we have shown that taking turn structure seriously makes the mathematical structure of ME games quite different from that of BM games. In particular, we've shown that several important properties of a winning play in a conversation have the structure of rhetorically decomposition sensitive goals, and hence are Π^0_2 complete. Since Π^0_2 complete conditions and beyond are in some sense truly infinitary, this technical result buttresses our philosophical claim that agents must plan their conversations as though they were infinite.

We have also provided a strong motivation for consistency rhetorical cooperativity, a key assumption of [2] even in the absence of other shared goals. And we have built these constraints into our conception of the Jury. We ex-

plored conversational goals which lie in the low levels of the Borel hierarchy with natural examples of conversations and provided a precise typology for them. We have shown how turns and turn involving winning conditions make ME games strictly more expressive than BM games and their winning strategies more complex. We also showed how intuitive constraints like consistency affect the complexity of winning conditions, moving many intuitive characterizations of winning conditions in conversation to Boolean combinations of Π^0_2 and Σ^0_2, or Muller, conditions in the Borel hierarchy. We have also provided an intuitive example of a Π^0_3 winning condition in ME games. We have also shown some linguistic consequences of this typology with respect to stability or instability of evaluation.

We believe we have just scratched the surface of a rich and interesting theory of the structure of conversations. There are many directions for further development. For instance, we have considered only very simple forms of our discourse constraints. For instance, let us consider consistency. Remaining consistent means avoiding a finite sequence of formulas in V in accordance with the logic that yields a proof of an inconsistency. The consistency predicate is Π^0_1 in the language of Peano arithmetic, since provability is expressible as a Σ^0_1 predicate. It is not definable in S1S (or equivalently LTL), and so the complexity of arithmetic consistency lies higher up in the Borel Hierarchy (or maybe even beyond that). We want to investigate more complex forms of other constraints with respect to the Borel Hierarchy and what this means for the existence and feasibility of winning strategies in ME games.

There are other directions for further research. One is to develop our model of the Jury further using techniques from mean-payoff games [48]. A major challenge for that is how to assign local per-move utilities to the players such that they converge to the required global utilities in the limit. Developing such a model would also give us a better and more sound way to talk about the finite satisfiability (refutability) of a winning condition by a play p. In mean-payoff games there are natural notions of 'optimal strategies' for the players, the 'value' of the game and 'quantitative determinacy'. We would then be able to prove things like: given certain kinds of winning conditions, if the Jury assigns a winner after k turns then it can be wrong by at most ϵ.

Another is to endow our ME games with epistemic elements by associating type-spaces for the players and belief functions for each type. In addition there are types for the Jury. The players start with an initial set of beliefs and dynamically revise their beliefs about the types of the other players and that of the Jury after every move. Using such a rich model, we

can explain many interesting phenomena – why players play the way they do and why they say what they say.

Finally, we believe that integrating signaling games with ME games is an exciting future area of study: each conversational move is in effect the result of a signaling game, and the ME game as a whole yields a sequence of iterated signaling games whose utilities are guided by the overall winning conditions.

6 Appendix

Proof of Proposition 1:

Proof Let $s(\cdot|m)$ be a receiver strategy. Let $\mu(\cdot|m)$ be a probability distribution over sender types such that s is rational given belief in μ and such that $s(a^*_{t_{\text{GOOD}},m}|m) > 0$. By definition, if s is rational given μ, $a^*_{t_{\text{GOOD}},m}$ must be a best response to m. In particular $a^*_{t_{\text{GOOD}},m}$ must yield a better (or equal) expected utility that $a^*_{t_{\text{BAD}},m}$ which can be written as:

$$\sum_{t\in T} \mu(t|m)u_R(t,m,a^*_{t_{\text{GOOD}},m}) - \sum_{t} \mu(t|m)u_R(t,m,a^*_{t_{\text{BAD}},m}) \geq 0$$

that is to say

$$\left(\begin{array}{l} \sum\limits_{t\in T_{\text{GOOD}}} \mu(t|m)\left((u_R(t,m,a^*_{t_{\text{GOOD}},m}) - u_R(t,m,a^*_{t_{\text{BAD}},m})\right) \\ -\sum\limits_{t\in T_{\text{BAD}}} \mu(t|m)\left(u_R(t,m,a^*_{t_{\text{BAD}},m}) - u_R(t,m,a^*_{t_{\text{GOOD}},m})\right) \end{array}\right) \geq 0$$

Notice that both terms of the above difference are positive (the first term on the left is strictly positive since $t_{\text{GOOD}} \in T_{\text{GOOD}}$). Let

$$\begin{aligned} \delta_{\text{GOOD}} &= max_{t\in T_{\text{GOOD}}}(u_R(t,m,a^*_{t_{\text{GOOD}},m}) - u_R(t,m,a^*_{t_{\text{BAD}},m})) \text{ and} \\ \delta_{t_{\text{BAD}}} &= u_R(t_{\text{BAD}},m,a^*_{t_{\text{BAD}},m}) - u_R(t_{\text{BAD}},m,a^*_{t_{\text{GOOD}},m}). \end{aligned}$$

Since $t_{\text{BAD}} \in T_{\text{BAD}}$ we have:

$$\sum_{t\in T_{\text{GOOD}}} \mu(t|m)\delta_{\text{GOOD}} - \mu(t_{\text{BAD}}|m)\delta_{t_{\text{BAD}}} \geq \left(\begin{array}{l} \sum\limits_{t\in T_{\text{GOOD}}} \mu(t|m)\left((u_R(t,m,a^*_{t_{\text{GOOD}},m}) - u_R(t,m,a^*_{t_{\text{BAD}},m})\right) \\ -\sum\limits_{t\in T_{\text{BAD}}} \mu(t|m)\left(u_R(t,m,a^*_{t_{\text{BAD}},m}) - u_R(t,m,a^*_{t_{\text{GOOD}},m})\right) \end{array}\right)$$

Hence we must have

$$\sum_{t\in T_{\text{GOOD}}} \mu(t|m)\delta_{\text{GOOD}} - \mu(t_{\text{BAD}}|m)\delta_{t_{\text{BAD}}} = \mu(T_{\text{GOOD}}|m)\delta_{\text{GOOD}} - \mu(t_{\text{BAD}}|m)\delta_{t_{\text{BAD}}} \geq 0$$

and since $\delta_{t_{\text{BAD}}} > 0$ by hypothesis we have $\mu(T_{\text{GOOD}}|m)\frac{\delta_{\text{GOOD}}}{\delta_{t_{\text{BAD}}}} \geq \mu(t_{\text{BAD}}|m)$ which concludes the proof. □

References

[1] N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

[2] N. Asher and A. Lascarides. Strategic conversation. *Semantics and Pragmatics*, Vol 6.2:1–62, 2013.

[3] N. Asher, S. Paul, and A. Venant. Conversational goals and achieving them in no-win situations. Submitted to *Journal of Logic Language and Information*, 2015.

[4] R. J. Aumann and S. Hart. Long cheap talk. *Econometrica*, 71(6):1619–1660, 2003.

[5] R. J. Aumann and M. Maschler. *Repeated games with incomplete information*. MIT press, 1995.

[6] R. M. Axelrod. *The evolution of cooperation*. Basic books, 2006.

[7] A. Benz, G. Jäger, and R. van Rooij, editors. *Game Theory and Pragmatics*. Palgrave Macmillan, 2005.

[8] J.R. Büch and L.H. Landweber. Solving sequential conditions by finite-state strategies,. *Transactions of the American Mathematical Society*, 138:367–378, 1969.

[9] C.C. Chang and H. J. Keisler. *Model Theory*. North Holland Publishing, 1973.

[10] K. Chaterjee. Concurrent games with tail objectives. *Theoretical Computer Science*, 388:181–198, July 2007.

[11] V. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

[12] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

[13] J. Farrell. Meaning and credibility in cheap-talk games. *Games and Economic Behaviour*, 5:514–531, 1993.

[14] M. Franke. Meaning and inference in case of conflict. In Kata Balogh, editor, *Proceedings of the 13th ESSLLI Student Session*, pages 65–74, 2008.

[15] M. Franke. *Signal to Act: Game Theory in Pragmatics*. ILLC dissertation series. Institute for Logic, Language and Computation, 2009.

[16] M. Franke, T. de Jager, and R. van Rooij. Relevance in cooperation and conflict. *Journal of Logic and Language*, 2009.

[17] J. Glazer and A. Rubinstein. Debates and decisions: On a rationale of argumentation rules. *Games and Economic Behavior*, 36(2):158–173, 2001.

[18] J. Glazer and A. Rubinstein. On optimal rules of persuasion. *Econometrica*, 72(6):119–123, 2004.

[19] E. Grädel. Banach-Mazur games on graphs. In R. Hariharan, M. Mukund, and V. Vinay, editors, *Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, pages 364–382, 2008.

[20] H. P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics Volume 3: Speech Acts*, pages 41–58. Academic Press, 1975.

[21] B. Grosz and C. Sidner. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204, 1986.

[22] B. J. Grosz and S. Kraus. Collaborative plans for group activities. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 367–373, Los Altos, California, 1993. Morgan Kaufmann.

[23] A. Kechris. *Classical descriptive set theory*. Springer-Verlag, New York, 1995.

[24] L. Lamport. Sometime is sometimes not never: On the temporal logic of programs. In *Proceedings of the 7th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 174–185. ACM, 1980.

[25] D. Lewis. *Convention: A Philosophical Study*. Harvard University Press, 1969.

[26] L. Libkin. *Elements of finite model theory*. Springer, 2004.

[27] P. Malone. *The life you save:Nine Steps to Finding the Best Medical Careand Avoiding the Worst*. Da Capo Lifelong, 2009.

[28] W. C. Mann and S. A. Thompson. Rhetorical structure theory: A framework for the analysis of texts. *International Pragmatics Association Papers in Pragmatics*, 1:79–105, 1987.

[29] D. A. Martin. Borel determinacy. *Annals of Mathematics*, 102(2):363–371, 1975.

[30] R. Mauldin, editor. *The Scottish Book. Mathematics from the Scottish Café*. Birkhaüser, 1981.

[31] R. McNaughton and S. Papert. Counter-free automata. In *Research Monograph*, volume 65. MIT Press, 1971.

[32] J. Oxtoby. The Banach-Mazur game and Banach category theorem. *Contribution to the Theory of Games*, 3:159–163, 1957.

[33] P. Parikh. Communication and strategic inference. *Linguistics and Philosophy*, 14(5):473–514, 1991.

[34] P. Parikh. Communication, meaning and interpretation. *Linguistics and Philosophy*, 25:185–212, 2000.

[35] P. Parikh. *The Use of Language*. CSLI Publications, Stanford, California, 2001.

[36] D. Perrin and J. E. Pin. *Infinite Words: Automata, Semigroups, Logic and Games*. Elsevier Publications, February 2004.

[37] M. Rabin. Communication between rational agents. *Journal of Economic Theory*, 51:144–170, 1990.

[38] J. P. Revalski. The Banach-Mazur game: history and recent developments. Technical report, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 2003-2004.

[39] H. Sacks. *Lectures on Conversation*. Blackwell Publishers, Oxford, 1992. Edited by Gail Jefferson. This is the published version of lecture notes from 1967–1972.

[40] L.M. Solan and P.M. Tiersma. *Speaking of Crime: The Language of Criminal Justice*. University of Chicago Press, Chicago, IL, 2005.

[41] A. M. Spence. Job market signaling. *Journal of Economics*, 87(3):355374, 1973.

[42] D. Traum and J. Allen. Discourse obligations in dialogue processing. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL94)*, pages 1–8, Las Cruces, New Mexico, 1994.

[43] R. van Rooij. Being polite is a handicap: towards a game theoretical analysis of polite linguistic behavior. In *TARK*, pages 45–58, 2003.

[44] R. van Rooij. Signalling games select horn strategies. *Linguistics and Philosophy*, 27:493–527, 2004.

[45] A. Venant and N. Asher. Ok or not ok? commitments, acknowledgments and corrections. In *Proceedings of Semantics and Linguistic Theory (SALT 25)*, Stanford, 2015.

[46] A. Venant, N. Asher, and C. Dégremont. Credibility and its attacks. In *Proceedings of Semdial 2014*, Edinburgh, Scotland, September 2014. Semdial.

[47] D. N. Walton. *LOGICAL DIALOGUE-GAMES*. University Press of America, Lanham, Maryland, 1984.

[48] U. Zwick and M. S. Paterson. The complexity of mean payoff games. In *Computing and Combinatorics*, pages 1–10. Springer, 1995.