Background information in conversation games

Lochlan Morrissey

Preprint (version 1.0: 2018-09-19).
Find the latest version at https://lmorrissey.info/.

Abstract

I propose a game theoretic model of conversations with two interlocutors and multiple turns. The model positions *background information* as the primary object of players' reasoning, so that the construction, management, and interpretation of background information is considered interlocutors' chief concern. Background information is understood as being composed of three sorts of information: (*a*) world knowledge, the players' beliefs about which possible worlds are candidates for the actual world; (*b*) context, a ledger of the utterances that have been produced in the course of the game, and secondarily the players' private information about their intentions and their interpretations; and (*c*) common ground, the information that is agreed upon for the purposes of the conversation (after Stalnaker 1975, 2002). I propose that, of these three, only common ground requires the definition of a first-class entity in the game model, and that world knowledge and context may be modelled by leveraging player beliefs and private histories, respectively. To model the utterance semantics, the common ground, and player interpretations, I use a subset of Discourse Representation Theory (Kamp, Genabith, and Reyle 2011; Kamp and Reyle 1993) that provides a fully-specified, first-order language. I introduce a new means of capturing interlocutor preferences, through the definition of *heuristics*, which are informally-defined expectations that interlocutors mutually share of one another that can be implemented formally in the game model. I illustrate how heuristics can be used, and highlight that a heuristics-based approach permits a flexibility and sensitivity to player goals that is not available with approaches that rely solely on solution concepts.

1 Introduction

Game models of linguistic interaction offer particularly useful insights into the interrelation of utterance design and utterance interpretation. Even in situations where players are opposed (see, for instance Asher, Paul, and Venant 2017; McCready 2015), intelligible communication requires that: (i) the Speaker be mindful of how the Hearer is likely to interpret her utterance; and (ii) the Hearer anticipate the Speaker's expectations, in order to understand her motives for producing that utterance at that time. Let us call this feature of game models *bidirectionality of reasoning*, since each player reasons about the reasoning of the other, and this informs both players' strategy selection. The bidirectionality is present in all game models of interaction, since, as I show below, it arises from the definition of strategic reasoning itself. Where game models of interaction differ is largely in what the primary subject of the players' bidirectional reasoning is. For instance, Parikh (2006, 2007) proposes a model in which interlocutors reason about the relation of speech acts to the situation that they are performed in; for Benz (2012a,b), interlocutors reason about whether their speech acts are free of errors. Others propose models in which the

reasoning itself is the subject of reasoning: Franke (2009, 2011) explicitly models player beliefs about the cognitive capacities of their adversaries; while McCready (2015) models players' perception of their adversaries' reliability.

In this paper, I propose a game theoretic model that positions *background information* as the primary object of players' bidirectional reasoning. This term is used in general to denote many sorts of information available to the players: for instance, the location and time of the exchange; the history of the interlocutors' relationship; and the interlocutors' beliefs about the world that they inhabit. Certainly, information that may be termed 'background information' has been included in previous game theoretic models. For instance, Parikh's (2006) model, mentioned above, invokes *situations* to contain this kind of information. But this is only the most explicit treatment: all game models of natural language conversation assume background knowledge of some sort, which is usually encoded into the Speaker's private information, called her *type*. The kind of background information that I am interested in for the purposes of this paper is the necessary and sufficient information that an agent requires when designing or interpreting an utterance. I identify three sorts of information that conforms to this goal: *world knowledge*—an agent's knowlege of the facts of the possible world that he considers true—, *contextual information*—knowledge of what has occurring in the conversation—, and *common ground propositions*—information that has been agreed upon for the purpose of the conversation (following Stalnaker 1975, 2002). In the model that I propose, each of these is represented explicitly and formally using a different game theoretic mechanism, meaning that the actual contents of the background information for each player is accounted for transparently. For instance, interlocutors' common ground is modelled explicitly by a semantic structure in a logical language based on Discourse Representation Theory (Kamp, Genabith, and Reyle 2011; Kamp and Reyle 1993).

The shift of focus to an examination of background information necessitates a means of modelling a conversation constituted of *multiple* turns. Since contextual knowledge and common ground propositions take multiple turns to come into being, the model must be one with multiple plays of a (modified) signalling game. However, the model I propose is not an iterated game model like those proposed by McCready (2015) and Asher, Paul, and Venant (2017), for players' utilities may change at each turn. Indeed, the motivation for proposing a new approach that treats background information as a first-class element of the game model arises from the study of conversations that involve multiple turns with contributions of multiple interlocutors. An example of a datum that is of interest is the following from Morrissey (2017, p. 199), which is the beginning of a scene of improvised theatre:

(1)
1. A: Well, they fired me.
2. B: Why'd you show up for work then, dad?
3. A: I've got nowhere else to go, do I? [*He mimes unpacking an unidentified object.*] Been here for … forty years. This is all I know.
4. B: Well, you've pulled out the old typewriter again!
5. A: [*He mimes typing on a typewriter.*] A job doing … A job worth doing's worth doing right, son.
6. B: [*Pause.*] Just turning on my PC.
7. A: Mine's already warmed up.
8. B: Dad, I … dad, I'm really worried about you, you know. I'm going to—look, I see you as a bit of a father figure. … You know, in—in kind of an abstract sort of

way … and the fact that … that—I mean, this is desperately *sad*. You'd—maybe you should—you've got nowhere else to go?

9. A: I've got nowhere else I want to be.
10. B: Than next to me?
11. A: Than at *work*.

Due to the spontaneous nature of improvised theatre—that is, given that nothing about the plot is determined prior to its execution—the interlocutors' background information when the scene begins is empty. This datum thus illustrates the coconstitution of the background information of the conversation. Consider, for instance, the second line, which establishes an absurd situation: that A is at work despite his having been fired. The line is a joke—but it only functions that way because the background information of the scene includes the assertion made by A's utterance in line 1, namely that he has been fired. We see a similar use of the background information as a resource against which interpretations are made in line 8—in which a joke ("I see you as a bit of a father figure … in an abstract sort of way") is made of A being B's father, which is established in lines 2 and 5. A more striking example in this datum is the graduation of the object that A unpacks from an unidentified object (in the performance, the actor literally mimes unpacking a box, and pulling out a nondescript object) to a typewriter. Note that it is only *after* B identifies the object as a typewriter that A begins to use it as such. B's utterance in line 4 asserts that the object is a typewriter. This becomes a part of the scene's background information, and is used as the basis of subsequent communication.

Data of improvised performance exhibit frequent, strong effects of background information on the subsequent interpretation of utterances. However, it appears that that these effects are not constrained merely to improvised performance, but are a feature of any conversation involving multiple turns and a stable set of interlocutors; see, for instance, empirical studies of conversation including Haugh 2010—which documents speakers' use of jocular mockery based on shared background information—, and Grainger, Mills, and Sibanda 2010—which investigates the role of culturally-mediated background information on communicative strategies.

The motivation in this paper is to provide a *general*, *descriptive* model of longer conversations. Since background information appears to be present in any given longer conversation, it provides a solid basis for the subject of interlocutors' bidirectional reasoning in a general model. This model is not intended to replace existing models of longer conversations (for instance, Asher, Paul, and Venant 2017), but instead aims to be able to be adapted for use with these models and those developed in the future. For this reason, the model that I introduce here is descriptive, rather than predictive. Similar to a game model without solution concepts, I identify a set of parameters whose specification produce a game model of longer conversations without producing a prescription of which strategies are selected by players. This does not mean that the model lacks expressive power; indeed, in section 5, I illustrate how the model is better equipped in certain ways to model longer conversations by defining multiple, interoperative *heuristics* than models with singular, monolithic solution concepts. That is, while solution concepts precisely identify an agent's optimal actions given certain circumstances, the heuristics that I propose model principles that interlocutors use to guide their decision-making. They may be thought of, then, as *hermeneutic principles*, and are used to include the effects of an interlocutor's experience, biases, and cultural mediation into the formal model.

The plan of the paper is as follows. In section 2, I introduce the game theoretic formalism that forms the basis of the model. In section 3, I specify what is meant by background

information, and in particular I identify three sorts of background information that I include in the model. I define the game model, its linguistic component, and the three sorts of background information in section 4. In section 5, I introduce heuristics, provide some examples thereof, and propose how they may be incorporated into a game model of linguistic interaction. The paper concludes in section 6.

2 Interpretation games

This section provides the definition of the basis for a (pseudo)iterated game model of communication. In section 2.1, I define a standard signalling model, and in section 2.2 I identify some characteristics of iterated games that are used in the model that I propose. In section 2.3, I describe two existing approaches to background information in game models.

2.1 Signalling games

The normal approach to modelling natural language conversation in game theory, including those in Benz 2012a; Franke 2011, is to modify a *signalling game* so that the signals are natural language utterances; that is, they have a natural language semantics. Signalling games take the following form:

(i) an impartial, disinterested player called Nature selects a state of the world $w \in W$[1] at random, with probability $\Pr(w)$;

(ii) only Speaker S observes the w that Nature selects, and sends a signal $m \in M$ to the Hearer;

(iii) Hearer H observes the m and selects an action $a \in A$.

The state of the world that is encoded in the game model is meant to broadly represent any information that only the Speaker has access to. This means that the state of the world may represent basically anything: the Speaker's opinion of cauliflower, yesterday's weather, the meaning of life, and so on. But the state always encodes what the Speaker intends to convey (or obscure, in uncooperative games) by her communicative act. Despite this model's simplicity, then, it captures a sort of bidirectionality of intentionality. In one direction, the Speaker designs her utterance so that a particular state w is conveyed to or obscured from the Hearer. The Speaker expects that the Hearer will interpret her utterance in a particular way, and she selects her strategy with this expectation in mind. These expectations are captured by the *Hearer's prior belief* ρ_H, which is a probability distribution over the set of states $\rho(W)$. An element of ρ_H is denoted $\Pr_H(w)$. The Hearer's prior belief expresses how probable the Hearer considers certain states to be *before she observes m*. This information can be mobilised by the Speaker to design her utterance in such a way that it leverages her estimate of the Hearer's expectations to design her utterances to render certain interpretations more salient. In the other direction, the Hearer's action gives the Speaker evidence of how the latter has interpreted the former's utterance by responding to the state that she considers most likely. Formally, H will select the action with the

[1] This is often referred to as the Speaker's *type* in game theoretic sources (see, for instance, Benz, Jäger, and Rooij 2006; Weibull 1997). I prefer the term *state of the world* only because it incorporates better with the intensional semantics that I introduce in section 4.1.1; as far as my model is concerned, there is no difference between these concepts.

highest *expected utility*. The expected utility for player i given play of a signalling game $\mathrm{s} = (w, m, a)$ is calculated by solving

$$\hat{u}_i(\mathrm{s}) \equiv \sum_{w \in W} \Pr(w) \times u_i(\mathrm{s}). \tag{1}$$

H's action demonstrates which state she believes is the true state, which is informed by how she interprets S's utterance.

The model of signalling games was devised by Lewis (1969) to describe abstract interaction that wasn't necessarily linguistic. To model linguistic interaction, let us add a mechanism for natural language semantics. The mechanism that is most widely used is an extensional model; a model in which each utterance is associated with its extension, that is a set of states in which that utterance is true. The resultant class of games is *interpretation games*, and an example of such a game is given by

$$\langle N = \{S, H\}, W, \rho_H, \sigma = W \times M \times A, [\![\cdot]\!], U \rangle, \tag{2}$$

where N is the set of players; S, H are the Speaker and Hearer, respectively; W is a set of states of the world; ρ_H is the Hearer's prior belief; σ is a set of plays of the game, such that each play is a vector of the form $\mathrm{s} = (w \in W, m \in M, a \in A)$; $[\![\cdot]\!] : M \longrightarrow W$ is a semantic denotation function that maps natural language sentences to the states in which they are true; and $U : \sigma \longrightarrow \mathbb{R}^2$ is a utility function that maps plays of the game to a 2-dimensional vector of real numbers, which represents the utilities for the pair (S, H).

In many game models of interaction (see, for instance, Benz and Rooij 2007; Parikh 1992), the players' utilities are assumed to be equivalent. This means that both S and H succeed if certain conditions are met, and they both fail if not. In interaction games, these conditions are specifications that certain Hearer strategies respond to certain states. Consider a simple game model in which $W = \{t_1, \ldots, t_x\}$, $A = \{a_1, \ldots, a_y\}$, and $x = y$ (that is, W and A are of equivalent size). A utility function is a *symmetric Lewis utility function* iff the matrix $W \times A$ is the identity matrix

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}. \tag{3}$$

That is, if $i = j$ then $u(w_i, m, a_j) = 1$ *for all* $m \in M$. Game models that employ this utility function treat the message instrumentally. That is, the message's actual content doesn't matter, and what matters is that it causes the Hearer to select a particular action a. The message is a conduit, therefore, between observation of w and a.

Finally for this exposition of the standard model of interpretation games, there are two additional belief structures that are not specified explicitly in the game model. The Hearer has two stages of belief, the first of which has already been discussed. The second arises after H observes m, called her *posterior belief*, which describes states are most likely given the evidence provided by receiving m. This is likewise a conditional probability distribution $\rho(W|M)$. The second of these additional belief structures is the *Speaker's belief*, which is a probability distribution $\rho(A)$. Since the Speaker knows that the Hearer will update her belief over W given the evidence, she selects an m that will cause the true value of w to be the maximal value for $\Pr(w|m)$. Likewise, the Speaker's belief alters her strategy selection to allow for what she expects the Hearer's response will be.

2.2 Iterated games

Recent work in formal linguistics (Asher, Paul, and Venant 2017; McCready 2015)[2] has adapted game theoretic formalisms of *iterated games* to model linguistic interaction. Iterated games are constituted of repetitions of a single *stage* game, called periods, where the periods usually involve each player of the game making a single move (see Abreu 1988; Fudenberg and Maskin 1986; Mertens 1986 for diverse applications of iterated games in economics). Importantly, the stage game is stable across these iterations, meaning that the set of players, the strategy set, and the utility function don't change throughout the iterated game.

Iterated games provide a useful theoretical basis for the model that I propose in this paper. However, their inflexibility with respect to utility functions means that they struggle to account for any shift in interlocutors' motives that might occur in the course of a conversation. (This is a core concern of the model that I propose here, and I discuss it further in section 5.) In my model, however, I do make use of two aspects of iterated game models:

1. Iterated games have *histories*, which are registers of the strategies that have been played during the periods of the iterated game. Consider the case of an iterated signalling game. Since $\sigma = W \times M \times A$, a history of an iterated game at period x, h^x, is a vector in σ^{x-1}; that is, σ to the $(x-1)$th power,

$$h = \underbrace{\sigma \times \cdots \times \sigma}_{x-1}. \tag{4}$$

   Histories are useful because they can be used to express the effects that strategies played in a period can have on later periods. Similar to beliefs, histories can be implemented directly into the deliberative process of the players. For instance, as per Mertens (1986), strategies can be defined as functions over a player i's action set A_i:

$$s_i(x) = \begin{cases} a \in A_i & \text{if } x = 1 \\ h^{t-1} \longrightarrow A_i & \text{if } x > 1 \end{cases} \tag{5}$$

   That is, in period 1, i's action is simply an action $a \in A_i$. Meanwhile, in later periods, i's choice is determined by a function from the history of the turns that precede period x onto A_i.

   In games of incomplete information, such as signalling games, players each have their own *private* history (Aumann, Maschler, and Stearns 1995). Private histories represent the actions that a particular player is able to observe. So, for instance, the Speaker in a signalling game can see the world selected by Nature, her own utterance, and the action selected by the Hearer; but the Hearer cannot see the first of these.

2. The length of iterated games does not need to be definite. Games of indeterminate length are called *infinitely iterated games*. Paradoxically, such games may have a finite length, but the length of the game isn't known by the players and the game may terminate after any given turn. According to Aaronson (2013), this lack of knowledge can have profound effects on the players' strategy selections.

[2] I describe these approaches below.

Games of indeterminate length are appropriate for longer conversations, since it is rarely (if ever) the case that interlocutors know in advance how many turns will constitute a conversation. And, as we will see, histories are used in my model to represent players' knowledge of what has been uttered in the course of a conversation.

2.3 Existing approaches to background information

Background information in game models tends to be represented as an emergent property of an interaction that arises either through players' linguistic reasoning (in equilibrium-based approaches) or through the sequential performance of utterances (in history-based approaches).

An equilibrium-based model of background information is exemplified by the treatment of scalar implicature proposed by Franke (2011), which I discuss here using my own vignette. Suppose that H wishes to know whether there are any bananas left at home. Her housemate, S, saw how much their other housemate ate last night: she either ate some but not all of the bananas ($w_{\exists\neg\forall}$), or else she ate all of the bananas ($w_{\forall}$). Let us suppose that the Hearer believes that these states have an equal probability of $1/2$. H asks how much bananas S ate, to which S can respond with one of two utterances:

(2) She ate some bananas last night. ($m_{\exists\neg\forall}$)

(3) She ate all the bananas last night. ($m_{\forall}$)

In response, H can either buy more bananas (a_+), or she can not buy any bananas (a_-). We assume a Lewis utility function such that the matrix $(w_{\exists\neg\forall}, w_{\forall}) \times (a_-, a_+)$ is an identity matrix

$$\begin{array}{c} \\ a_- \\ a_+ \end{array} \begin{array}{c} \begin{array}{cc} w_{\exists\neg\forall} & w_{\forall} \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}. \tag{6}$$

The complication in this example arises in the semantic denotation function: (2) is true whether the players' housemate ate all the bananas or not, while (3) is true iff the housemate ate all of the available bananas. So, $[\![m_{\exists\neg\forall}]\!] = \{w_{\exists\neg\forall}, w_{\forall}\}$ and $[\![m_{\forall}]\!] = \{w_{\forall}\}$.

The equilibrium point becomes clear when we consider that Hearer will play a_+ if she observes $m_{\forall}$, because $\Pr(w_{\forall}|m_{\forall}) = 1$ while $\Pr(w_{\forall}|m_{\exists\neg\forall}) = 0$. Although $m_{\exists\neg\forall}$ is ambiguous, for it may be uttered truthfully both if S observes $w_{\exists\neg\forall}$ and $w_{\forall}$ so that $\Pr(w_{\exists\neg\forall}|m_{\exists\neg\forall}) = \Pr(w_{\forall}|m_{\exists\neg\forall}) = 1/2$, it is the only available utterance whose extension contains $w_{\exists\neg\forall}$. Therefore, if the Speaker wishes to be optimally clear, she will select $m_{\forall}$ if $w_{\forall}$, which leaves $m_{\exists\neg\forall}$ if $w_{\exists\neg\forall}$. This approach leads both Franke (2009, 2011) and Rooij (2008) to claim that game models of communication directly model "linguistic contexts", and similar approaches to background information based on the same notion have been proposed by Parikh (2006, 2007), and by Benz (2012a,b).

A second approach to background information arises naturally from models of iterated games, through this isn't necessarily a primary feature of these models. For instance, the iterated model proposed by Asher, Paul, and Venant (2017) defines *Message Exchange (ME) games*, which are adapted Mazur-Banach games (see Soare 2016). The ME game model is specifically meant for ME games have two players, a strategy set constituted of strings for each player, and a winning condition for each player that specifies a set of strategies such that if a player plays a strategy in this set, the game ends; since ME games are of indeterminate length, they are infinitely iterated. An innovative feature of the ME game model is the presence of a Jury, which is "an abstract scoring device" (Asher, Paul, and Venant 2017, p. 376) that enforces certain cooperative practices between antagonistic agents. More

saliently, the Jury determines whether new information introduced is consistent with old information, and in this way acts as a device that judges whether new information is able to be accommodated with existing contextual information; this is, in essence, what the common ground does in my model. Another instance of a similar approach is proposed by McCready (2015), who uses histories that record what players have observed, and allow them to make judgments of other players' motives. In particular, McCready's model describes how reliability arises between interlocutors, and how the past actions of an interlocutor affect how their fellow interlocutors behave. Histories play a similar role in my model in providing a ledger of the actions that have been performed during a conversation and informing future behaviour.

3 Three sorts of background information

As I mentioned in the introduction, I propose that the background information that an interlocutor has access to at any given turn can be modelled by specifying three interrelated entities: the interlocutor's world knowledge; the conversation's context; and the common ground shared by all parties to the conversation. In the model that I propose here, these are together considered the necessary and sufficient epistemological conditions for a participant of a conversation to make a meaningful contribution to that conversation. They are necessary because without access to these three pieces of information, a Speaker is unable to reason about how the Hearer is likely to interpret her utterance, and a Hearer lacks the ability to anticipate the Speaker's reasoning. Indeed, I claim that these three sorts of background information provide the minimally sufficient amount of linguistic data that interlocutors must possess to facilitate successful communication in longer conversations. World knowledge and the context have both been identified in previous studies as necessary for speaker interaction (see, for instance, Benz 2012a,b; Parikh 2006, 2007). Common ground, meanwhile, adds a shared semantic resource that underpins a conversation, and provides interlocutors with a source of information that is used when arriving at interpretations. I discuss this more below. This means that a specification of these three entities is not exhaustive with respect to background information, and it is intended that they be use as the basis for further game theoretic study of longer conversations; I illustrate in section 5 that heuristics can be used to add additional interlocutor expectations to the model.

With this in mind, the following subsections specify the scope of each of these entities, and how each is implemented in the game model.[3]

3.1 World knowledge

An interlocutor's world knowledge is the incomplete image that she has of the possible world that the interaction takes place in. The inclusion of world knowledge as a necessary constituent of background information is motivated by insights gained from prior game theoretic research, in which the state of the world plays an important part in motivating certain speech and interpretation acts. As such, I model this sort of knowledge as a set of possible worlds that she considers plausible *candidates* for the actual world, which are the worlds consistent with what the interlocutor knows and believes about the actual world. This indirect means of modelling world knowledge is based on the following intuition. Since any interlocutor's world knowledge doesn't include a complete specification of any

[3]However, formal definitions for each of these entities are provided in section 4.

given possible world, it is is a *partial model* of a that world, and it can therefore be 'embedded' within a hypothetical, complete specification of the possible world. Given that there are potentially infinite possible worlds, an interlocutor's world knowledge will partially model a large number of possible worlds. Therefore, modelling the interlocutor's world knowledge as a set of candidate worlds achieves a similar effect to modelling the interlocutor's knowledge directly using, for instance, formulae. This means of modelling world knowledge is chosen because it integrates well with game models of player belief. As I highlighted in section 2, beliefs provide a means of modelling how players expect their fellows to behave, the effects that players expect their actions to have, and ultimately how these relate to the possible world of the interaction. I will also illustrate in section 4.1.2 that beliefs can be used to directly model the effects of the common ground on the players' cognition.

3.2 Context

The context, as I use it in this paper, is somewhat different to usual understandings of context, as in Rooij 2008, Franke 2011, and Parikh 2006. In these sources, context is associated with background information in general, and contextual information provides a basis for the production and reception of given speech acts in a conversation. My usage of context is somewhat more specific than this: the context is a register of the utterances that have been produced in the course of a particular conversation. Naturally, the register is immutable, insofar as it is impossible to delete utterances that have already been produced in a conversation. The scope of the context is 'local', in that it does not contain any background information that is not introduced directly during the conversation, and it includes all information that is introduced during a conversation, *even if that information is self-contradictory*; two utterances that respectively entail p and $\neg p$ may both be included in a single context. The context is included as a constituent of background information to explicitly model the effects that arise from a series of utterances, and particularly how these effects influence interlocutors' decisions at a particular turn. The context is modelled by the game's history, and especially by players' private histories. Histories are especially useful in this context because they allow for the formal implementation of effects of decisions on subsequent turns. This allows us to model the way that interlocutors alter their linguistic decisions based on the direction of the conversation.

3.3 Common ground

The common ground is the most complex of the three kinds of knowledge, and its introduction is the chief contribution of this paper to the game theoretic study of linguistic interaction. It is based on Stalnaker's (1975, p. 273) notion of common ground as the information which an interlocutor "can use as a resource for the communication of further information, and against which he will expect his speech acts to be understood". This notion of common ground is, I argue, equivalent to the commonplace understanding of context, and in later work, Stalnaker (2002, p. 716) specifies an additional constraint of *acceptance* for common ground propositions: "it is common ground that φ in a group if all members accept (for the purpose of the conversation) that φ, and all believe that all accept that φ, and all believe that all believe that all accept that φ, etc.". So, a group's common ground is a set of propositions that are accepted *for the purposes of the conversation*, without requiring that any of the group actually commit to the truth of these propositions. It is this property that allows the common ground to act, as Stalnaker claims, as a resource against which interlocutors expect their utterances to be understood; the common ground

acts as an agreed-upon 'semantic lingua franca' that permits the successful interpretation of utterances during a conversation. A simple example of how common ground affects interpretation is in the assignment of a pronoun to a particular individual. It is agreed upon by the interlocutors that, for a certain period of a conversation, an individual will be denoted by *she* and *her*, and it is only through agreeing to revise the denotation of these terms that they can shift to apply to another individual.

Defining the common ground thus positions *negotiation* at the centre of communication. A basic kind of negotiation is implied by the bidirectionality of reasoning that I identified above: in cooperative games, each player's decision is taken by considering her fellow's motives, and by adjusting her deliberation to take account of these motives. This is not *explicit* negotiation, but it still involves compromise, consideration of the fellow's motives and reasoning, and so on. By including common ground, however, a more radical idea of negotiation emerges. In the course of a conversation, multiple interlocutors contribute information that might be accepted into the common ground, or they contest information that is already present in the common ground. The common ground's constitution is thereby negotiated by the parties to a conversation. Since the common ground is required to successfully interpret utterances, and since common ground propositions are necessarily agreed upon, the process of interpreting an utterance relies on negotiation.

A negotiated conception of interpretation underpins a significant amount of research on communication, including general theoretical approaches proposed by Arundale (2006, 2008, 2010), Cooren and Sanders (2002), Goffman (1955, 1967), Hymes (1962, 1963), and Silverstein (1993, 2001), and recent studies in face theory (Bargiela-Chiappini 2003; Bargiela-Chiappini and Haugh 2009; Spencer-Oatey 2007). The negotiated nature of interpretations is captured particularly well by Arundale (2010, p. 2080), and it is worth quoting him at length, for the sequence that he describes is readily applicable to common ground propositions (emphasis is Arundale's):

> As [an interlocutor] Amy designs and produces her first position utterance, she anticipates [a second interlocutor] Bob's interpreting of it. Her utterance, like any other, affords and constrains the possibilities for interpreting meaning and conversational action. Because she is cognitively autonomous from Bob, because meanings are not determinate (Arundale 2008, p. 244), and because (unlike an analyst) she cannot know what Bob's second position utterance will be, Amy does not know just how Bob has interpreted her utterance until she interprets his second position utterance responsive to it [...] Prior to that point, Amy's projecting of Bob's interpreting of her first position utterance remains *provisional*, because she does not know how it will be understood *within the frame of their particular interaction* [...] In producing his second position utterance, Bob provides evidence of some aspects of his interpreting of Amy's first utterance. Her interpreting of that evidence allows her either to confirm her provisional projecting if that interpreting appears consistent with her projecting, or to modify her provisional projecting if she finds inconsistency. At the point she has interpreted Bob's second position utterance, Amy's projecting of Bob's interpreting of her first utterance, or her modification of it, becomes an operative interpreting because she now has evidence for how it has been understood within their particular conversation. If Amy is to add a third position utterance, she must in some manner take this operative interpreting into account, whether or not it is consistent with her own initial projecting. A provisional interpreting, then, is one not yet assessed in view of uptake, even though one may be certain about it, while an operative interpreting is one assessed in view of uptake, even though one might change it.

There are two things about this process that I wish to emphasise. First, a range of interpretations is projected by Alice's utterance, and, while she may have a specific interpretation in mind, Bob can select *any* interpretation, *even if that interpretation was not considered*

possible by Alice. I argue that this suits a game model of interpretation, since it means that the set of possible interpretations is infinitely large, and that the primary factor in selecting an interpretation is the players' beliefs. This removes the need to artificially restrict the possible interpretations of an utterance, resulting in a more general, robust model of interpretation. Second, following from this first, utterances that follow an initial utterance m render particular interpretations of m more plausible. That is, there is a cumulative effect exerted by the addition of information during a conversation.

I claim that this process equally applies to the establishment of common ground propositions. First, a proposition is introduced by an interlocutor i through the production of an utterance. This proposition affords a range of interpretations, one of which is selected by the second interlocutor j. j's interpretation informs the selection of a second utterance, which entails an acceptance, partial acceptance, or rejection of the proposition. If it is accepted, it enters the common ground, and if it is disputed, then it does not. It is through the iterations of this process of introduction and judgement that a common ground is built, and the process is subsequently affected by the ever-changing common ground. I discuss this process in greater detail in section 4.4.3.

How I propose to model common ground is as a first-class entity of the game model definition that captures information agreed upon by the players of the game. To this end, I use a subset of Discourse Representation Theory to define a simple model of context that is primarily meant to capture the cumulation of information during a conversation, and that is, importantly, specifically designed to be incorporated into game theoretic models of interaction. For instance, this means that my model of context uses the turn-taking structure of a game model instead of defining information states, as is the case in more comprehensive models of context (see Groenendijk and Stokhof 1991; Veltman 1996). Discourse Representation Theory provides a useful basis for a model of context because it was designed to represent multi-utterance conversations. But is also useful because Discourse Representation Structures—the basic unit of analysis in Discourse Representation Theory—can represent both individual utterances and partial representations of possible worlds. This ability derives from their simplicity. DRSs are constituted of two sets: a set of entities, and a set of conditions on these entities. They can therefore represent each of the three levels of knowledge that I identified earlier in the paper: an interlocutor's world-knowledge, the context of the conversation, and the accepted common ground of the conversation. I define this formal language in the next section.

4 The game model

Let us now turn to a complete statement of the model of interaction that obtains from introducing background information as the primary object of players' reasoning. Formally, the game model is a tuple:

$$\langle N, \mathcal{L}, \mathfrak{M}_{\mathcal{L}} = \langle W, \mathfrak{D}, \mathfrak{I} \rangle, \rho_S, S = W \times M \times K, \Gamma, [\![\cdot]\!], U \rangle, \tag{7}$$

where $N = \{i, j\}$ is a set of players; $\mathcal{L}$ is a DRS language; $\mathfrak{M}_{\mathcal{L}} = \langle W, \mathfrak{D}, \mathfrak{I} \rangle$ is a model for $\mathcal{L}$; ρ_S is the Speaker's belief for that particular turn; σ is the strategy space of the game; M is a set of natural language utterances; K is a set of valid interpretations, defined as setences of $\mathcal{L}$; Γ is the set of all valid common grounds; $[\![\cdot]\!]$ is a semantics function; and U is a utility function.

N, σ, $[\![\cdot]\!]$, and U are each defined in the same way as the standard game theoretic definition (see section 2). On the other hand, $\mathcal{L}$, $\mathfrak{M}_{\mathcal{L}} = \langle W, \mathfrak{D}, \mathfrak{I} \rangle$, and Γ—which form the linguistic component of the model—require definition, and ρ_S requires a redefinition

based on the introduction of the common ground. I therefore introduce the linguistic model in section 4.1, and the redefinition of player belief in section 4.2, before returning to describe some properties of this game model at the level of individual turns (section 4.3) and multiple turns (section 4.4).

4.1 The linguistic model

In this section, I define the formal language based on DRT that will be used as the semantic component of the game model (section 4.1.1), and I discuss the formal implementation of the common ground, and some properties thereof (section 4.1.2).

4.1.1 The formal language

As I stated above, the formal language that I use to model context and common ground is a subset of Discourse Representation Theory, which is defined in the following paragraphs, following the specification given by Kamp, Genabith, and Reyle (2011).

Syntax Let $\mathcal{L}$ be a first-order language with a vocabulary consisting of: Ref, a set of discourse referents; Name, a set of one-place definite relation constants; a set Π^n of n-ary predicate constants; and a set $\{=, \neg, \wedge, \vee, \forall, \exists\}$ of logical symbols. Sentences of $\mathcal{L}$ are defined inductively:

(i) if $D \subseteq \mathrm{Ref}$ (called the sentence's *universe of entities*), and C is a set of valid conditions (defined below), then $k = \langle D, C \rangle$ is a sentence;

(ii) if $x, y \in \mathrm{Ref}$, then $x = y$ is a valid condition;

(iii) if $a \in \mathrm{Name}$ and $y \in \mathrm{Ref}$, then $a(y)$ is a valid condition;

(iv) if $\pi \in \Pi^n$ is an n-place predicate, and if $x_1, \ldots, x_n \in \mathrm{Ref}$, then $\pi(x_1, \ldots, x_n)$ is a valid condition;

(v) if k is a sentence, then $\neg k$ is a sentence;

(vi) if k and k' are sentences, then $k \wedge k'$, $k \vee k'$, and $k \implies k'$ are sentences; and

(vii) if k is a sentence, then it is a valid condition.

The *merge* of two sentences k and k', $k \uplus k'$, is a sentence made by the pairwise union of (a) the universe of each of the sentences, and (b) the set of conditions of each of the sentences:

$$\langle D_k \cup D_{k'}, C_k \cup C_{k'} \rangle. \tag{8}$$

Note that condition (vii) allows for recursive embedding of sentences within sets of conditions. This condition in particular makes the DRT-based language extremely useful for representing emerging discourses.

Sentences serve a few different functions in my model. First, they operate as formulae of the first-order language $\mathcal{L}$, and are treated in the same way as such formulae are in the model proposed by, for instance, Benz (2012a). Second, with further specifications, they may function as contexts and common grounds. I deal with these additional usages below.

Semantics The semantics of $\mathcal{L}$ are intensional, and it is primarily by this that the formal language is able to be incorporated into the game model. The intensional model $\mathfrak{M}_{\mathcal{L}}$ includes a nonempty set of worlds W, a nonempty domain of individuals $\mathfrak{D}$, and an interpretation function $\mathfrak{I}$. The interpretation function accepts two sorts of input: (i) names, for which it is defined $\mathfrak{I} : \mathrm{Name} \rightarrow \{\{d\} |\ d \in \mathfrak{D}\}$; and (ii) n-ary predicates, for which it is defined $\mathfrak{I} : \Pi^n \rightarrow (W \rightarrow \wp(\mathfrak{D}^n))$.

Recall that DRSs, from which the definition of sentences is derived, are meant to represent models of a discourse, and can therefore represent partial models of a possible world. This notion provides the basis for a definition of truth in DRT: if for each $x \in D_k$ there is a $y \in \mathfrak{D}$ such that every condition on x holds, then k holds. That is, if there is a *homomorphism* from D_k into $\mathfrak{D}$, then k holds. This is expressed formally as a *verifying embedding*. Let $\mathcal{L}$ be a DRS language, and let $\mathfrak{M}_{\mathcal{L}} = \langle W, \mathfrak{D}, \mathfrak{I}\rangle$ be an intensional model of $\mathcal{L}$. A verifying embedding f for a DRS (or sentence) k into $\mathfrak{M}$ for a (possibly empty) set of discourse referents $X \subseteq \mathrm{Ref}$ is a homomorphism from X into $\mathfrak{M}$, $f : X \rightarrow \mathfrak{D}$. Embeddings are defined at the level of the model, and not at the level of individual worlds; and, as such, they are understood as holding for all worlds. That is, a verifying embedding exists for a particular sentence at a particular world or it does not. Verifying embeddings represent *partial assignment* of variables to individuals, so that an embedding $f(x) = d$ associates a variable $x \in k$ with an individual $d \in \mathfrak{D}$; these assignments are called partial, since they are not exhaustive specifications of all individuals $\mathfrak{D}$. Note that if an assignment cannot be made—that is, if there is no d that x can be assigned to—then $x \notin \mathit{Dom}(f)$, where $\mathit{Dom}(f)$ is the domain of the embedding f. Multiple embeddings may be associated by an extension function, which is important to the formal definition of sentence truth. Let f and g be verifying embeddings. f *extends* g to the discourse referents X, denoted $f \subseteq_X g$, iff (i) $\mathit{Dom}(f) = \mathit{Dom}(g) \cup X$; and (ii) for all $x \in \mathit{Dom}(g)$, $f(x) = g(x)$.

On the basis of its assignment of variables to indiviudals, an embedding f may *verify* a condition c at a world w, denoted $f \models_{\mathfrak{M},w} c$. For each of the DRS conditions defined above, the following defines verification by f:

(i) $\langle f, g\rangle \models_{\mathfrak{M},w} \langle D, C\rangle$ iff $f \subseteq_D g$, and for all $c \in C$, $g \models_{\mathfrak{M},w} c$;

(ii) $g \models_{\mathfrak{M},w} (x_i = x_j)$ iff $f(x_i) = f(x_j)$;

(iii) $f \models_{\mathfrak{M},w} a(x)$ iff $\mathfrak{I}(a) = \{f(x)\}$ (where $a \in \mathrm{Name}$);

(iv) $f \models_{\mathfrak{M},w} \pi(x_1, \ldots, x_n)$ iff $\langle f(x_1), \ldots, f(x_n)\rangle \in \mathfrak{I}(\pi)$;

(v) $f \models_{\mathfrak{M},w} \neg k$ iff there is no g such that $\langle f, g\rangle \models_{\mathfrak{M},w} k$;

(vi) $f \models_{\mathfrak{M},w} k \vee k'$ iff there exists some g such that either $\langle f, g\rangle \models_{\mathfrak{M},w} k$, $\langle f, g\rangle \models_{\mathfrak{M},w} k'$, or both;

(vii) $f \models_{\mathfrak{M},w} k \implies k'$ iff for all g such that $\langle f, g\rangle \models_{\mathfrak{M},w} k$, there exists a h such that $\langle g, h\rangle \models_{\mathfrak{M},w} k'$.

Condition (i) of this list defines the truth of a sentence. If a sentence is true at world w, we may omit the embedding function and write simply $\models_{\mathfrak{M},w} k$, since embeddings hold for all worlds in W. A sentence's proposition is the set of worlds at which that sentence is true: $[\![k]\!] \equiv \{w \mid \models_{\mathfrak{M},w} k\}$.

The formal language defined here is only a subset of DRT, and it does not include more advanced features of the theory (see Kamp, Genabith, and Reyle 2011 for a very comprehensive overview). Yet, its ability to index individuals in a possible world and assign predicates to them, and the possibility of recursively embedding sentences allows for it to assume the multitude of functions that I identified above.

4.1.2 The common ground γ

The common ground is modelled by a sentence γ, whose entities and conditions specify the information that has been agreed upon by the parties to a conversation. Common grounds may recursively contain sentences in their conditions, and update of the common ground is represented by merging a sentence that expresses some information with an existing common ground. Unlike contexts, the common ground may not contain any contradictions. It is simple to see why this is the case: the propositions p and $\neg p$ may not be agreed to simultaneously. If p is present in the common ground of a particular conversation, and if $\neg p$ is subsequently introduced, then further negotiation must take place to decide which of these is agreed to. This condition is expressed formally as follows:

Definition 4.1. *A sentence γ is a valid common ground iff $[\![\gamma]\!]$ is nonempty.*

This definition excludes those sentences that contain contradictions, since for any valid sentence k, $[\![k \wedge \neg k]\!] = \{\}$. There are no other conditions on common grounds.

Since common grounds are sentences, truth of a common ground γ at a particular world w is determined precisely as truth of a sentence: if a pair of embedding funcions exist for γ at w, then $\mathfrak{M}$ models γ at w; $\models_{\mathfrak{M},w} \gamma$. Recall that this entails that *all* the conditions of the common ground must be met for it to be true at a particular world. Common grounds are by definition subjective structures that exist only in the minds of the interlocutors, and therefore truth may not appear to be of limited concern. However, truth allows common grounds to be associated with extensions, which play a major role in players' deliberation, as I demonstrate below. The definition of a common ground's extension is the same as a sentence's extension:

$$[\![\gamma]\!] \equiv \{w \mid \models_{\mathfrak{M},w} \gamma\}. \tag{9}$$

The most important function that the common ground fulfils is to *constrain possible interpretations*; this is the aspect of common ground identified by Stalnaker (1975, p. 273) that allows an interlocutor to use it as a resource "against which he will expect his speech acts to be understood". It does so by staking a semantic space of agreed-upon conditions that must be taken into consideration when an interpretation is selected. This notion is similar to the notion of semantic narrowing in relevance theory (Carston 2004; Sperber and Wilson 2004). As the common ground becomes more comprehensive, the range of possible interpretations diminishes. In fact, this characteristic of common grounds is able to be demonstrated formally:

Theorem 4.1. *As $|C_\gamma|$ increases, $|[\![\gamma]\!]|$ monotonically decreases.*

Proof. Consider two sentences k, k', and let $\mathrm{k} = k \uplus k'$. As per the definition of truth of a sentence k given, an embedding h must exist at world w such that for all $c \in C_k$, $h \models c$. Since $C_\mathrm{k} = C_k \cup C_{k'}$, and since a sentence's extension is the set of worlds at which that sentence is true, $[\![\mathrm{k}]\!]$ is the set of worlds in which the conditions in both C_k and $C_{k'}$ are true; that is $[\![\mathrm{k}]\!] = [\![k]\!] \cap [\![k']\!]$. And since, for any two arbitrary sets A, B, $|A \cap B| \leq \max(|A|, |B|)$, $|[\![\mathrm{k}]\!]|$ cannot be larger than the larger of $[\![k]\!]$ and $[\![k']\!]$. □

This means that as a common ground is established, certain possible worlds become more salient, and certain interpretations are rendered more probable.

4.2 Player beliefs with a common ground

Once a common ground has been established, the core mechanism by which it influences the deliberation of interlocutors is by acting on players' beliefs. Specifically, the definition of players' beliefs are redefined to be probability distributions that are conditioned on the common ground. Thus, the Speaker's belief is redefined from a probability distribution over K to $\rho_S(K|\Gamma)$. This expresses the Speaker's anticipation of the Hearer's behaviour given the common ground, and, as I noted above, it gives the Speaker a basis for predicting the Hearer's actions.

Likewise, the Hearer's beliefs are redefined to include common ground. Recall that the Hearer's *prior* belief models her judgment of the probability of each $w \in W$. Like the Speaker's belief, the Hearer's prior belief is updated to include the common ground as a conditional term:

$$\rho(W|\Gamma). \tag{10}$$

Once the Hearer observes the utterance, she updates her prior belief to incorporate the addition of information provided by the utterance. In the standard game model, the Hearer's posterior belief is modelled as a conditional probability distribution over W given m. When common ground is included, the hearer uses both the utterance *and* the common ground to judge which worlds are most likely. As such, there are two conditional terms in the Hearer's posterior belief:

$$\rho_H(W|M,\Gamma). \tag{11}$$

Following the definition provided by Russell and Norvig (2010), the probability of a particular $\Pr_H(w|m,\gamma) \in \rho_H(W|M,\Gamma)$ is found by solving

$$\Pr_H(w|m \cap \gamma) \equiv \frac{\Pr(m \cap \gamma|w) \times \Pr(w)}{\Pr(m \cap \gamma)}. \tag{12}$$

I have mentioned that player beliefs, so defined, represent their expectation of their fellows' behaviour. However, since these beliefs target the set of states, they also represent players' world knowledge, and beliefs about which states are candidates for the actual state. By incorporating common ground into their redefinition, a mechanism for common ground to affect world knowledge is provided. This means that, as a conversation unfolds, players update their world knowledge based on the composition of the common ground, and acceptance of facts into the common ground is likewise affected by players' world knowledge.

4.3 Individual turns

The game model treats a conversation as a sequence of *turns*, which are constituted of three actions:

1. Nature selects a possible world $w \in W$;
2. a player $i \in N$ produces an utterance $m \in M$; and
3. a player $j \in N$ such that $i \neq j$ selects an interpretation k.

Note that the indices i and j are used instead of S and H because players change roles at each turn. The player who produces an utterance in turn t selects an interpretation in $t+1$. In this paper, I only consider cases where there are two players, and so i is the speaker at every odd-numbered turn, while j is the speaker at every even-numbered turn. When I discuss players in particular turns, I will refer to them as the Speaker or Hearer of the turn.

As I stated above, conversations are modelled as having an indeterminate number of turns. At the conclusion of each turn, players are informed of two things: their private history, and their expected utility. A major difference between conversation games and regular interpretation games in that *both* players have private information. Only the Speaker is able to view Nature's selection of a possible world; and since interpretation is an action that occurs within the Hearer's mind, only she can observe which interpretation she has selected. So, the only move that is common to the histories of both players is the utterance that the Speaker produces. In my model of interaction, however, both the Speaker and the Hearer have *beliefs* about the private information of the other player.

Definition 4.2. *The Speaker's private history for a given turn h_S is $h_S = (w, m, \max \rho_S(K))$, and the Hearer's is belief $h_H = (\max \rho_H(W))$. Note that $\rho_S(K)$ is the Speaker's belief, and that $\rho_H(W)$ is the Hearer's posterior belief.*

The game's history, and therefore the formal expression of its context, is a register of *visible* actions, namely of utterances. It is given by the following definition.

Definition 4.3. *The context of a conversation at turn t is*

$$\bigcap_{i \in N} h_i^t, \tag{13}$$

and is therefore a member of M^{t-1}.

Players are informed of their expected utilities at the conclusion of each turn. Interpretation games being a game of incomplete information, the players view their *expected utilities*, which are found by solving

$$\hat{u}(w, s_i, s_{-i}) \equiv \sum_{w \in \rho_i} \Pr(w) \times u(w, s_i, s_{-i}). \tag{14}$$

Following Aumann, Maschler, and Stearns (1995), players are not informed of their actual utilities until the conclusion of the game. Hence at any given turn, players also have what I call a *provisional total utility*, which is the sum of the expected utilities of the proceeding turns. The provisional total utility function takes a history at turn t, h_i^t:

$$\hat{u}_i(h_i^t) \equiv \sum_{\tau \leq t}^{\tau=1} \sum_{w \in W} \Pr(w|h_i^\tau) \times u(h_i^\tau). \tag{15}$$

Since they model interlocutors' judgments of the success of the conversation, provisional total utilities are private knowledge of each player. Note that these judgments may differ drastically from the actual utility that a player is awarded at the conclusion of the game. Hence, the model captures potential errors in a player's estimation of her own performance. Given that both players' private histories are determined according to their respective beliefs, they may mistakenly guess that a particular world or interpretation has been selected, and thereby misjudge the felicitousness of their strategy. The establishment and management[4] of the common ground, then, becomes a means of *mitigating differences in players' estimated utilities* by providing them with evidence that certain interpretations have been selected. This becomes further apparent when we consider how the common ground behaves over multiple turns in the next section.

[4]See Krifka 2007 and Repp 2012 for a discussion of common ground management, and strategies that are used to deliberately and purposefully control the information that is in the common ground.

4.4 Conversations of multiple turns

To illustrate how this model of individual turns operates when iterated multiple times, consider the initial two turns of a conversation game. I will present two separate versions of this game that demonstrate both how background information accumulates for players of conversation games, and how the common ground in particular is affected by a process of proposal and subsequent acceptance or rejection. In both, i's utterance attempts to convey w. In the first game, j accepts i's proposal into the common ground. In the second game, j rejects it. For both of these examples, $N = \{i, j\}$, and i is the Speaker in turn 1, and j is in turn 2. We treat these two turns as a short segment of the overall game, so that the players expect the game to continue beyond turn 2.

4.4.1 Acceptance

Turn 1 A $w \in W$ is selected, and is observed by i before selecting her utterance $m_i \in M$. Observe that i's context (that is, her history) and the common ground are both empty; $h_i^1 = \emptyset$ and $\gamma^1 = \emptyset$. i's deliberation rests *chiefly on her world knowledge*; that is, her belief of j's interpretation given the common ground. Since the common ground is empty, however, j does not take it into account, and it is excluded from the formal model of her beliefs. i's deliberation is therefore identical to an agent in a single-play interpretation game. j observes m and selects an interpretation $k_j \in K$. Like i, j doesn't have access to a context, and the common ground term in j's belief is discounted, so her deliberation relies solely on her world knowledge.

In this turn, i proposes some information that j accepts in the next turn. Prior to j's verbal acceptance, however, she selects an interpretation that aligns with her own world knowledge, which causes her to accept the information. Let k_{m_i} be a DRS that represents m_i. Since w represents i's communicative intent, let us assume that $w \in [\![k_{m_i}]\!]$. If k_{m_i} is consistent with j's prior belief, then for each $w \in [\![k_{m_i}]\!]$, it is also the case that $\Pr_j(w) \in \rho_j(W)$ is nonzero. If m_i is an utterance that successfully conveys w, then j's interpretation—selected in light of her posterior belief in this turn $\rho_j(W|M, \emptyset)$—will have an extension that includes w.

At the conclusion of turn 1, each player is informed of her private utility. union Suppose that i believes that j will select an interpretation that is consistent with w, k_j. Her private history will reflect this: $h_i^1 = (w, m_i, k_j)$. Likewise, suppose that j's private history reflects that w is most probable: $h_j^1 = (w, m_i, k_j)$. Each player also estimates her utility, which is added to her provisional total utility.

Turn 2 Once a common ground has been established, the remainder of the turns more closely resemble one another. The players' decisions are made on the basis of the beliefs described in section 4.1.2; at each turn the common ground increases in specificity, reducing the number of worlds that are candidates for w. The decisions that are made later in the game therefore differ from those earlier, which conforms to an intuitive understanding of how utterances are selected and interpreted later in conversations.

As I have specified, j understood i's communicative intent, and, since her beliefs are aligned with what i proposes, j accepts i's proposal. j's communicative intent is to communicate her acceptance; I represent this by saying that Nature selects w' in this turn. j observes w' and produces an utterance m_j, such that $w' \in [\![k_{m_j}]\!]$. But since j is accepting w, $w \in [\![k_{m_j}]\!]$ as well. This means that the context of the conversation *and* the world knowledge *both inform the design of utterances*, and that these utterances are designed with a sensitivity to the contents of the common ground. The subsequent interpretation is also

sensitive to the three kinds of knowledge, as specified in the expression of the Hearer's posterior belief (see equation 12 above), since the common ground is populated by j's response.

4.4.2 Rejection

Turn 1 Like the example above, a $w \in W$ is selected, and is observed by i before selecting her utterance $m_i \in M$. Suppose that $w \in k_{m_i}$, and that $\Pr_j(w)$ is zero. j's interpretation of m_i will conclude that w is false—expressed by $w \notin [\![k_j]\!]$—and therefore that she should not accept the information into the common ground. Suppose that i believes that j believes that w, which leads her to misjudge j's interpretation, so that her private history is w, m_i, k'_j. Here we find that the model I propose here can account for errors in communication, and can model the potential effects of those errors on subsequent decisions.

Turn 2 j observes w', and produces an utterance m_j such that $w' \in k_{m_j}$ and $w \notin k_{m_j}$. Since j does not accept w, it is not included in the common ground of the conversation, and subsequent turns of the conversation operate without a common ground until one is established. The turn concludes in the usual way.

4.4.3 Cycles of negotiation

Although they are brief, these two examples demonstrate the centrality of negotiation to the formal model of common ground that I have sketched here. Common ground is unique amongst the sorts of background knowledge precisely because it is contested. Players' world knowledge is influenced only by inferences that arise on the basis of their fellows' actions. The context may be misinterpreted, yet there remain a 'true form' of the ledger of players' actions. The common ground, on the other hand, arises only through interlocutors' deliberate actions to bring it about and populate it. There are manifold means of managing the common ground, which can be distilled into three basic functions: proposal, acceptance, and rejection. The model that I propose, then, treats a conversation as a series of cycles of negotiation, and it posits that these cycles of negotiation are the primary level at which interlocutors participate in and construe a conversation.

Until now, I have been vague on how utility is determined. Instead of deciding on set criteria for the calculation of utility, I propose a means of establishing player preferences that permits preferences to change at each turn, and that allows for more complex accounts of player motives. I call these *heuristics*, and describe them in the next section.

5 Shifting motives and heuristics

I have argued throughout this paper that long conversations ought to be modelled as a series of interpretation games. Locating the most basic element of a conversation at the level of the individual speech act–interpretation selection dyad allows for a model that is sensitive to effects arising from subtle changes in the constitution of the interlocutors' context and common grounds. For what remains of the paper, I highlight another capacity that this model possess: the ability to model how an agent's motives shift in the course of a conversation. In game theoretic models of linguistic interaction, it is usually the case that players have an unwavering motive: their preferences over outcomes of the interaction do not change. This is trivially the case for singleton games, in which a player only has one opportunity to make a move, and therefore is not given the chance to alter her motives throughout. In the iterated models of linguistic interaction that we have already

seen, players generally have a single motive for selecting certain moves. This is clearest in noncoooperative games of linguistic interaction, most notably the models proposed by Asher, Paul, and Venant (2017) and McCready (2015), discussed above.

Yet even in strictly cooperative games, players' motives are dynamic. Consider another datum from Morrissey (2017, p. 144) of a scene of improvised performance set on the American frontier, evocative of spaghetti Western films.

(4)
1. A: What a whirlwind journey we've been on, Frank!
2. B: What bloody tumbleweeds I've seen, Jimmy!
3. A: I'm glad we came aw—
4. B: [*He begins to mime twirling a pistol at his hip.*]
5. A: —hey! Don't wave that thing at me!
6. B: Hey, I'm just twirlin' my gun.
7. A: Well twirl it out there, like everybody else. To the open *ranges*, Frank…
8. B: [*He turns away from A, towards the audience.*]
9. A: We came here for good reason.
10. B: [*turning back to face A*] It doesn't have the feel to it as it does when you twirl it in front of another *man*.
11. A: Don't! Frank!
12. B: Now I feel like I'm *twirlin'* for a *reason*, like I could just *draw* at any moment!
13. A: Frank! Comin' out west was a risk in itself. We left our families, a reliable food sauce, this is the last kind of risk I wanna take!
14. B: Sorry, just something about twirlin' a gun just makes me feel *alive*, man…
15. A: Nah, I get that. Back east wasn't the same.
16. B: No. It was frowned upon.
17. A: You couldn't even twirl a *toothpick* in your *mouth*.
18. B: No. They hated that.

The rest of the scene continues with the actors discussing twirling guns, and eventually ends with the players in a standoff, aiming their pistols at one another.

What's notable about this scene is that each of the actors appears to have her own preferred direction for the scene's plot. A continuously references the situation "back east", and compares it with the location of the scene, the western frontier (see lines 9 and 13, for instance). It appears that A intends for the scene to pertain in some way to their trip to the west, while B continues to reference the act of twirling pistols. The clearest manifestation of this disagreement is line 15: it seems that A means to reference the situation in the east being different, but in line 16, B recasts A's utterance to be a reference to *the act of twirling one's pistol in the east.* Not only does this illustrate that the meaning of A's utterance in line 16 arises through negotiation, it also suggests that A changes his preferred direction of the plot to conform to B's intention. While this alteration of motive may be captured using a 'monolithic' solution concept—for it may be the case that A's utility falls below an acceptable level as the result of the prolonged disagreement—I argue that allowing for players to change their preferences at each game without positing a singular, rigid definition for this change is more productive in longer conversation games.

In its present form, my model is able to capture these shifts in player motive. There are two reasons why this is the case. First, and as has already been discussed, the game model

of conversations is not a model of iterated games. This removes the requirement that each period of the game have precisely the same utility function. Since utility functions define a preference ordering over the space of possible outcomes, they implement a player's motivations, for it is according to an outcome's utility that a player selects her actions. And therefore, by not stipulating that a single utility function applies to all periods of a conversation game, the model can account for changes in motivation during the course of the game. Second, the model does not define equilibria, nor does it provide a means of deriving them. This follows naturally from the reason that I gave above: if the utility function changes continuously, then it is impossible to provide a single definition that will capture an equilibrium point for the entire conversation. The loss of equilibrium-based solution concepts might appear drastic at first, especially since many other game theoretic models of linguistic interaction rely on them (see, for instance Parikh 2001; Rooij 2008). Equilibrium solutions are used to provide a prediction of how agents will behave, which, importantly, are also used by *players themselves* to form expectations of how their fellows will behave. Since equilibria provide focal points for agent behaviour, they are useful in modelling players' long-term expectations of how their fellows will behave. However, the solution concepts that define the conditions of an equilibrium are necessarily rigid and are not amenable to conversation environments in which the players' motives change over time—in, for example, conversations constituted of cycles of negotiation.

To fill the gap left by a lack of solution concepts, I propose to include defined principles, called *heuristics*, that represent interlocutors' expectations of how their partners will behave. The heuristics themselves are *informal statements* that are associated with formal definitions that express these statements. They relate specifically to the *function* or *purpose* of a particular strategy, and specify conditions of optimality that relate to this function. Consider the following trivial example. Suppose that a Speaker at a particular world wishes to tell the truth at a certain turn. There is a set of utterances that are *optimally truthful*, namely the set of utterances that are verifiable at that particular world. A heuristic for optimal truthfulness would simply specify that any utterance in the set of verifiable utterances at a particular world is optimally truthful. Observe that the heuristic defines optimality according to a specific characteristic of the utterance, and so multiple heuristics may apply to a given utterance. Many of the utterances in the set of optimally truthful utterances may not be optimally relevant (according to some heuristic of optimal relevance), for instance. The application of multiple heuristics to utterances means that a set of heuristics can be defined so that, in aggregate, the heuristics represent the lens through which an interlocutor views the actions of her partners.

Heuristics in this sense have been used in a few game theoretic models, especially those that incorporate the Gricean maxims.[5] Rooij (2003), for instance, derives two 'rules of thumb' from the maxims. These are expressed formally as conditions of optimality referring to Speaker strategies: optimal relevance, and optimal coding (that, is optimal utterance design). Benz and Rooij (2007) likewise propose formal pragmatic principles, based on the maxims, that are distinct from solution concepts. Like these approaches, the heuristics that I propose for inclusion in my model apply only to Speaker strategies. That is, utterances can be optimal according to heuristics, but interpretations cannot. However,

[5]Grice's *cooperative principle* proposes that, in general, interlocutors expect that their fellows will adhere to the principle "make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (Grice 1989, p. 26). Grice (1989, p. 27) also proposes four maxims that are subordinate to the principle: 1. Quantity: make your contribution exactly as informative as is required, and no more; 2. Quality: "try to make your contribution one that is true"; 3. Relation: "be relevant"; 4. Manner: "be perspicuous". See Davies 2007 and Horn 2004 for further discussion of the principle and its maxims.

like other aspects of game theory such as beliefs and knowledge of prior moves, knowledge of the heuristics affects choices made both by the Sender and by the Receiver. The Speaker's choice of strategies is guided by their motives, which may cause her strategies to have a particular characteristic at a particular turn. If it is in her interests to play a strategy that is *optimal* according to a particular characteristic, then this choice is affected by the heuristics. The effects of the heuristics on the Receiver's strategy selection are subtler. They arise from the heuristics' role in the Receiver's inference of the intended function of an utterance by noticing how that utterance relates to the heuristics. If a heuristic defines a particular condition of optimality, then utterances that display characteristics consistent with this condition provide evidence to the Receiver that the utterance has been selected with the heuristic in mind. This evidence in turn provides grounds for the Receiver to infer the Sender's motive in producing an utterance.

5.1 Three heuristics

In this section, I give some examples of heuristics formulated by Morrissey (2017) that describe player expectations specifically in the context of improvised theatre. These are presented for a few reasons. First, they provide an idea of the general form of these heuristics. Second, they highlight the degree to which the heuristics may be bound to a specific genre of speech, or else that they may aspire—as Grice's cooperative principle does—to greater generality. Third, they demonstrate how informal statements may be formalised. The three heuristics that I present in these subsections describe optimal markedness (section 5.1.1), optimal convergence (section 5.1.2), optimal compatibility (section 5.1.3).

5.1.1 Optimal markedness

The first heuristic that I present refers to the status of potential interpretations of an utterance, and how certain interpretations become more salient as further information is added to the common ground. It describes a basic intuition that interlocutors appear to have that, in responding to an utterance, interlocutors implicitly indicate their interpretation of the utterance. When this is repeated with multiple interlocutors responding to a single utterance, then a particularly marked interpretation emerges. Morrissey (2017, p. 146) proposes the following heuristic to capture this: *Given a particular utterance, and a series of responses such that one response is made by each of the parties to a conversation, an interpretation of the utterance is optimally marked if it is most probable given* all *of the responses.* This statement is itself the heuristic, which is implemented in the formal model by specifying a criterion for optimality, given in definition 5.1.

Definition 5.1. *Let G^{n+1} be a conversation game constituted of $n+1$ games (where n is the number of players), such that g^1 has a specific Sender i, and such that for the following games $g^2, \ldots, g^n$, each player in N (including i) is the Sender for exactly one game. Let s^t denote a strategy profile of length t. An interpretation $k \in K$ is* optimally marked *if it solves*

$$\arg\max_{k \in K} \Pr(k|\{m|m \in \mathrm{s}^{n+1}\}). \tag{16}$$

The situation that is modelled by this heuristic is the following: i produces an utterance, which is interpreted by all other players. Each player, including i, produces another utterance, which is interpreted by the other players. At the end, a certain interpretation of i's initial utterance emerges as the most salient interpretation. The applicability of this to general conversation is clear, especially in light of Arundale's (2010, p. 2080) schema for the interpretation of an utterance quoted above.

5.1.2 Optimal convergence

The second heuristic also relates to the establishment of an operational interpretation, but, instead of describing a situation with multiple interlocutors, it describes a series of utterances. The general idea is this: if a series of utterances, taken together, denotes only w, then that series *converges* on w. The statement of this heuristic given by Morrissey (2017, p. 148) is: *Given an intended interpretation and a series of utterances meant to convey this interpretation, a strategy of repetition achieves optimal convergence if and only if the intersection of the extensions of each of the utterances contains* only *the intended interpretation.* Formally, this is defined for a series of utterances that are produced across a number of turns of the conversation game, as per definition 5.2.

Definition 5.2. *Let G^t be a conversation game constituted of t turns. A series of utterances $\langle m^1, \ldots, m^t \rangle$, where each is produced in its own period, is* optimally convergent as a repetition strategy *with respect to a particular $w \in W$ iff (1) $w \in [\![k_{m^1} \uplus \cdots \uplus k_{m^k}]\!]$; and (2) there is no $w' \in w$, such that if $w' \neq w$, then $w' \in [\![k_{m^1} \uplus \cdots \uplus k_{m^k}]\!]$.*

This heuristic affects both the design and the interpretation of a series of utterances. The effect on interpretation is perhaps less clear than that of the design. Recall the bidirectionality of reasoning inherent in game theoretic approaches. Interpretation of an utterance strategy involves the Hearer guessing the motive for the Speaker having produced that utterance at that time. This is also true of series of utterances, and if the Hearer notices that the Speaker's utterances are designed in such a way that their cumulative effect is to emphasise a $w \in W$, then—provided her preferences align with such an interpretation—she will select an interpretation that is consistent with w.

5.1.3 Optimal compatibility

The final heuristic that I include here pertains specifically to the common ground. It proposes a criterion that specifies which utterances are compatible with what has been established in the common ground of a conversation. The statement of this heuristic is simple: *A strategy is optimally compatible if the facts expressed by that strategy do not contradict what is common ground prior to that strategy's execution.* The formal expression of this heuristic is also simple. An optimally compatible strategy is one whose DRS representation has an extension that is a subset of the common ground's extension. That is, the DRS representation is true for each world that the common ground is.

Definition 5.3. *Let γ^t be a common ground at turn t. The set of optimally compatible strategies S^* is the smallest set such that, for each $s^* \in S^*$, $[\![k_{s^*}]\!] \subseteq [\![\gamma^t]\!]$.*

This heuristic may appear trivial at first, since it is obvious that an utterance doesn't contradict the common ground, then it is compatible with the common ground. Recall, however, that the expression of interlocutor expectations, even if they appear obvious, allows for greater insights into interlocutor behaviour, particularly in instances in which the interlocutors contravene these norms of behaviour.

5.2 Operationalising heuristics in the absence of solution concepts

I have argued that, in conversation games, a set of heuristics is preferable to singular utility functions and solution concepts that are defined for the entire exchange because the former allow for the greater degree of flexibility that is required when discussing conversations of multiple turns. During conversations, interlocutors' motives shift, depending

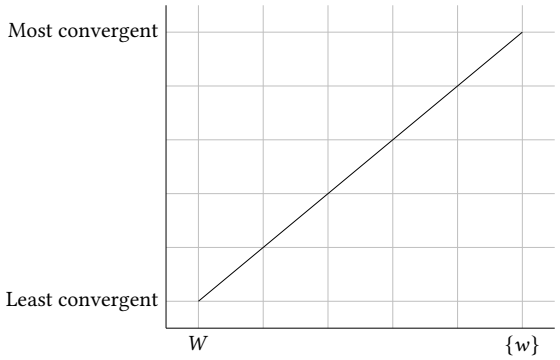

Figure 1: Illustration of the linear relationship between the convergence of an utterance series with respect to state w and the probability of w. The x-axis represents two extremes of $[\![k_{\mathrm{m}}]\!]$: at its optimum, $[\![k_{\mathrm{m}}]\!] = \{w\}$, and at its pessimum, $[\![k_{\mathrm{m}}]\!] = W$ (that is, the entire set of possible worlds).

primarily on what I have identified as background information: the direction of the conversation (that is, context), alterations to the speakers' world knowledge, and construction and management of the common ground. In this section, I briefly illustrate how heuristics can be employed directly to (*a*) derive utility over multiple turns; (*b*) derive utility within a single turn; and (*c*) elucidate interlocutors' cognition by highlighting potential effects that emerge from the operation of *multiple heuristics simultaneously*.

To do so, let us consider a game in which players use the heuristic for optimal convergence that I defined in section 5.1.2. Recall that a series of utterances $\mathrm{m} = \langle m^1, \ldots, m^k \rangle$ is optimally convergent with respect to w if this state is the only state in $[\![k_{\mathrm{m}}]\!]$—that is, in $[\![k_{m^1}, \ldots, k_{m^k}]\!]$. Since this heuristic's condition of optimality defines an upper bound to a strategy's convergence, derived from the number of states in $[\![k_{\mathrm{m}}]\!]$, there are series of utterances that are less convergent than others. It appears intuitive to claim that the more convergent a strategy is with respect to a w, the greater the probability of that w. We might imagine that such a relationship is linear, as in figure 1. Note that this implies that the convergence of a series of utterances with respect to w can be measured using $\Pr(w|\mathrm{m})$. I will use this fact in what follows.

Suppose that there are two players, S and H. For the sake of introduction, let us assume that S only produces utterances and H only selects interpretations—that is, the game models a delivered speech.[6] The epistemological conditions of games with background information still apply: players may only view their private histories, and beliefs are conditioned by the common ground. I assume for simplicity's sake, too, that there are two states w and w' with respective probabilities $\Pr(w)$ and $1 - \Pr(w)$. Imagine that S wishes to persuade H that w is the true state of the world, and that there are two hypothetical series of ten utterances, m and m', that are viable strategies. Figure 2 illustrates how each of these might converge with $\Pr(w)$ and $\Pr(w')$. In this example, w is equally likely given both m' and m for the first five turns. But after turn six, m''s convergence with respect to w declines, and the opposite holds for m. If the Speaker wishes to convey w, clearly m is the better series of utterances, and m' better conveys w'. This immediately suggests a preference-ordering for the two series of utterances, and thereby allows us to derive a

[6] This is technically possible with the model that I have proposed here, provided that an empty utterance is included in the set of utterances. The empty utterance is performed when interlocutors refrain from producing an utterance when it is their turn. Under this view, H forgoes her allotted turn to speak, and S selects an interpretation at the end of this turn.

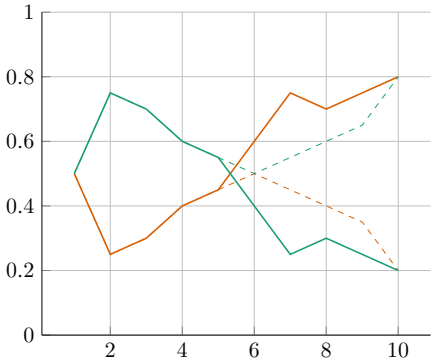

Figure 2: $\Pr(w|m)$ (orange) and $\Pr(w'|m)$ (green) over ten turns. The solid and dashed lines represent m and m', respectively.

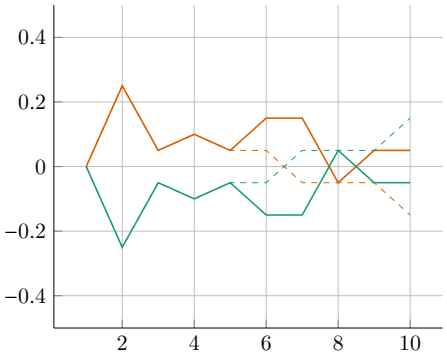

Figure 3: Marginal probabilities $\Pr(w|m)$ (solid line) and $\Pr(w'|m)$ (dashed line) for m (orange) and m' (green) over ten turns.

crude utility function, which happens to be a Lewis utility function:

$$\begin{array}{c} \\ \mathrm{m} \\ \mathrm{m}' \end{array} \begin{array}{c} \begin{array}{cc} w & w' \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}. \tag{17}$$

But even this simple example also allows us to define a less trivial utility function that can apply to individual utterances. When Speakers utilise convergence strategies, they are motivated by conveying the state that the utterances converge on. That is, they prefer utterances that contribute to the convergence to those that detract from it. A utility function that models players' preference for convergence strategies may, I propose, use the rate of change of $\Pr(w)$ between turns. Figure 3 obtains from plotting the marginal probabilities at each of the turns of these utterance series. The use of marginal probabilities to define the utility function permits an ordering of possible utterances from those that contribute least to the convergence of a series with respect to a particular state to those that contribute most. If a player intends to produce a convergent series, then her preferences will align with this ordering. And so, the heuristic can be used as the basis for a utility function. Indeed, such a method of calculating a player's utility can be incorporated into the model that I have proposed here, since the expected utility of a move, and the provisional total utility of the game, may both be calculated directly from marginal probabilities.[7]

[7] I do not discuss potential effects that heuristics have on expected and total provisional utilities here—but I expect that this would provide fertile ground for future research.

I have argued throughout this paper that heuristics are preferable to singular solution concepts because they capture interlocutors' nuanced motives, and how these can change in the course of a conversation. The model must, therefore, be able to account for the effects of multiple heuristics on interlocutors' decisions. I propose that a simple method of doing so is to calculate utility as a summation of weighted indices, where the indices each represent the degree to which a strategy is optimal according to a heuristic's criterion, and where the weights represent the relative importance that the interlocutor places on that particular heuristic's criterion. By defining weights for each criterion, we obtain a method of *ordering heuristics themselves*, which intuitively appears salient to interlocutors' deliberations. At particular points in the conversation, an interlocutor may wish to produce utterances that are optimally convergent (as in the examples above), while convergent strategies may be detrimental at other points of the game. To provide another example, consider adherence to Grice's maxims. At one point of a conversation, being optimally relevant may be preferred by an interlocutor, while flouting this optimality criterion—and giving rise to implicature—may be preferred at another point. Indeed, it is conceivable that an utterance might fulfil multiple optimality criteria at once, and that the interaction of these criteria will exert a strong effect on players' choices.

Formally, the implementation of criterion-weighting is as follows. Let O be a set of optimality criteria (that is, formally-defined heuristics), and let $o_i(s)$ be the score that a strategy s is awarded according to criterion $o_i \in \mathcal{H}$. s be a strategy profile, and V be a set of weights such that the number of weights is the same as the number of optimality criteria, and such that $\sum_{v \in V} v = 1$. A utility function U is *heuristics-sensitive* if it is calculated by solving

$$u_S(\mathrm{s}) = \sum_{i=1}^{|O|} v_i o_i(s). \tag{18}$$

By specifying that the sum of the weights must equal 1, and by thereby forcing weights to have values, the model requires that players must account for all of the criteria, and that they cannot be undecided on the importance of a particular criterion. I believe this is a more credible model of how interlocutors weigh up the importance of a number of different factors than other models that do not place such restrictions.

Criterion weighting is a simple method for incorporating players' deliberation of heuristics themselves, yet it provides nuanced understandings of interlocutor decision making. In particular, it allow us to identify *thresholds of minumum acceptability for potential strategies* given a certain configuration of weights. To demonstrate this, let us suppose that there are two heuristics—inspired by Grice's (1989, p. 27) injunctions to "be perspicuous", to "make your contribution as informative as required", and to not "make your contribution more informative than is required"—of optimal clarity and optimal efficiency, which are respectively encoded by criteria o_x and o_y. Imagine that an increase in the optimality in one sees a decrease in the optimality of the other, but that this negative correlation isn't precise, and describes a tendency: utterances that are optimally clear tend to be inefficient, since they tend to be exhaustive, and utterances that are optimally efficient tend to be unclear, since they tend to be too concise.[8] A player who is more interested in producing clear utterances is prepared to sacrifice the efficiency of her utterance, and *vice versa*: as a result of the weighting of clarity over efficiency, losses in efficiency affect the utility of the expected utterance less than gains in clarity. The configuration of weighting determines how much of her preferred criterion the player is willing to sacrifice.

[8] This is an extremely simplified example of a problem that has been well-studied in the linguistics literature; see Rooij 2003, Benz 2012a,b, and Franke 2011.

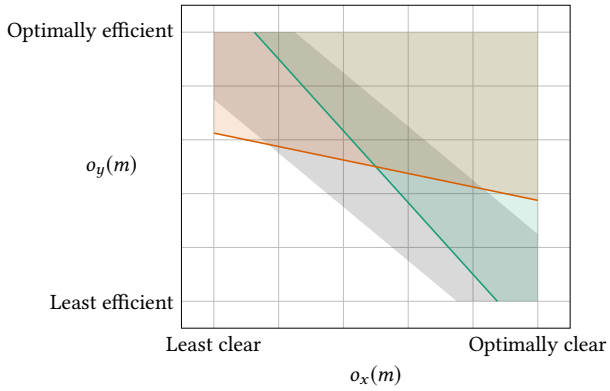

Figure 4: Acceptable utterances for players with weighting configurations $\{2/3, 1/3\}$ (green) and $\{1/8, 7/8\}$ (orange), bounded by minimum acceptability thresholds. The gray shaded area represents realisable utterances.

Consider two example weighting configurations for o_x, o_y: $V = \{2/3, 1/3\}$, and a more extreme $V = \{1/8, 7/8\}$. The spaces of acceptable utterances for each of these, which are bounded by the minimal acceptable thresholds plotted in figure 4.[9] The green shaded area contains all realisable utterances under these two optimality criteria (since an increase in one correlates with a decrease in the other—though recall that this relationship is not exact). The green and orange shaded areas represent acceptable utterances for those weighting profiles, and the intersection of these areas with the green area contains the utterances that are both acceptable and realisable. It is these utterances that the player with that weighting configuration will select from.

Note that one can't derive optimal *strategies* from this information, since strategies are optimal not only according to the weighting of heuristics, but rely on sensitivity to the other components of the model that I have specified. But these example weighting configurations do illustrate that, just by formally defining optimal *criteria* and by assigning weights to these criteria, we can identify which strategies a player considers acceptable. Furthermore, weighting configurations can be altered at each turn, so a player's optimal strategies can shift at each turn. This approach, then, can capture the complex process of deliberation that Speaker undergoes when assessing which linguistic strategy to select. It may also be mobilised to model a Hearer's deliberation, for if the Hearer is aware of the Speaker's weighting configuration, then she will interpret the utterance according to the heuristics. The heuristic-sensitive utility function that I have introduced in this section may, therefore, capture the bidirectionality of reasoning, and may therefore function as a viable alternative to standard, equilibrium-based solution concepts.

6 Conclusion

The model that I have proposed in this paper presents background information as the primary object of players' bidirectional reasoning. It specifies three sorts of information that interlocutors reason about: world knowledge, context, and common ground. The last of these is the only sort of background information that I propose be included as a first-class

[9] These curves are found by plotting the straight line

$$(-1 + v_y - v_x)x + (1/2 + v_y), \tag{19}$$

where v_x is the weight of criterion o_x.

entity of the game model. The other two have instead been modelled by leveraging standard components of game models, namely player beliefs for world knowledge, and public and private histories for context. The linguistic component of the game model is captured by a subset of Discourse Representation Theory, which performs many representational roles in the model, namely the implementations of the common ground, of utterance semantics, and of interpretation strategies. Since background information arises over time (particularly context and common ground), the games defined by the model have an indeterminate number of turns. Conversation games, however, are not strictly iterated games. While they rely on some existing research on iterated games—notably the use of private histories to model context, they are not iterated because utility functions may change. It is assumed in this model that player motives shift throughout a conversation, and therefore that a fixed utility function that applies to every turn of the game is not appropriate. This introduces an instability to the game model that renders assigning individual solution concepts at best spurious, for calculating equilibria in such an environment is barely feasible. And so instead, I introduce heuristics into the model, which represent principles according to which players expect their fellows to behave. These heuristics are informally-stated principles that are implemented formally in the game model, similarly in style to game theoretic implementations of the Gricean maxims.

I have intended for this model to be formulated simply enough that it may be used as the basis for further models of long conversations. By leveraging of existing components of the game model to capture world knowledge and context, these sorts of background information can be readily included in other game models. Furthermore, the use a first-order language to define common ground means that any game model with a first-order language can incorporate common ground, substituting the DRT-based language used here for the analyst's preferred language. Most important, I think, is the introduction of heuristics to a game model of natural language interaction, which allows for the model to be tailored to specific genres. Following Morrissey (2017), I have proposed heuristics that apply specifically to speech in improvised theatre. It is likely that other genres of speech (such as interviews, debates, naturally-occurring speech between two speakers who don't share a common mothertongue, and so on) impose expectations on the behaviour of their participant interlocutors. Indeed, it is clear that each communicative culture operates with its own set of heuristics.[10] The definition of heuristics provides a precise method of capturing qualities of agents' cognition, and provides a grounding for their deliberation that is sensitive to the communicative context. Further work developing the use of heuristics in different genres, with a greater number of interlocutors, and with different motives (especially those in which the players' utilities are disaligned) would deepen the precision of game models of natural language interaction, and would provide a more solid basis for the analysis of linguistic strategy.

References

Aaronson, Scott (2013). "Why philosophers should care about computational complexity". In: *Computability: Turing, Gödel, Church, and beyond*. Ed. by B Jack Copeland, Carl J Posy, and Oron Shagrir. Cambridge: MIT Press.

Abreu, Dilip (1988). "On the theory of infinitely repeated games with discounting". In: *Econometrica* 56.2, pp. 383–396.

[10] For indicative research from the vast body of work on politeness studies, see Bargiela-Chiappini and Kádár 2010 and Haugh 2004, 2007.

Arundale, Robert (2006). "Face as relational and interactional: A communication framework for research on face, facework, and politeness". In: *Journal of Politeness Research* 2.2, pp. 193–216.
— (2008). "Against (Gricean) intentions at the heart of human interaction". In: *Intercultural Pragmatics* 5.2, pp. 229–258.
— (2010). "Constituting face in conversation: Face, facework, and interactional achievement". In: *Journal of Pragmatics* 42.8, pp. 2078–2105.
Asher, Nicholas, Soumya Paul, and Antoine Venant (2017). "Message Exchange Games in Strategic Contexts". In: *Journal of Philosophical Logic* 46.4, pp. 355–404.
Aumann, Robert J, Michael Maschler, and Richard E Stearns (1995). *Repeated games with incomplete information*. Cambridge: MIT Press.
Bargiela-Chiappini, Francesca (2003). "Face and politeness: New (insights) for old (concepts)". In: *Journal of Pragmatics* 35.10, pp. 1453–1469.
Bargiela-Chiappini, Francesca and Michael Haugh (2009). *Face, communication and social interaction*. London: Equinox Publishing.
Bargiela-Chiappini, Francesca and Dániel Z Kádár (2010). *Politeness across cultures*. Basingstoke: Palgrave Macmillan.
Benz, Anton (2012a). "Errors in pragmatics". In: *Journal of Logic, Language and Information* 21.1, pp. 97–116.
— (2012b). "Implicature of complex sentences in error". In: *Practical Theories and Empirical Practice: A linguistic perspective*. Ed. by Andrea C Schalley. Amsterdam: John Benjamins, pp. 273–306.
Benz, Anton, Gerhard Jäger, and Robert van Rooij (2006). "An introduction to game theory for linguists". In: *Game Theory and Pragmatics*. Ed. by Anton Benz, Gerhard Jäger, and Robert van Rooij. Basingstoke: Palgrave Macmillan, pp. 1–82.
Benz, Anton and Robert van Rooij (2007). "Optimal assertions, and what they implicate. A uniform game theoretic approach". In: *Topoi* 26.1, pp. 63–78.
Carston, Robyn (2004). "Relevance theory and the saying/implicating distinction". In: *Handbook of Pragmatics*. Ed. by Laurence R Horn and Gregory Ward. Oxford: Blackwell, pp. 633–656.
Cooren, François and Robert E Sanders (2002). "Implicatures: A schematic approach". In: *Journal of Pragmatics* 34.8, pp. 1045–1067.
Davies, Bethan L (2007). "Grice's cooperative principle: Meaning and rationality". In: *Journal of Pragmatics* 39.12, pp. 2308–2331.
Franke, Michael (2009). "Signal to act: Game theory in pragmatics". PhD thesis. Universiteit van Amsterdam.
— (2011). "Quantity implicatures, exhaustive interpretation, and rational conversation". In: *Semantics and Pragmatics* 4.1, pp. 1–82.
Fudenberg, Drew and Eric Maskin (1986). "The folk theorem in repeated games with discounting or with incomplete information". In: *Econometrica* 54.3, pp. 533–554.
Goffman, Erving (1955). "On face-work: An analysis of ritual elements in social interaction". In: *Psychiatry* 18, pp. 213–231.
— (1967). *Interaction ritual: Essays on face-to-face behavior*. New York: Anchor Books.
Grainger, Karen, Sara Mills, and Mandla Sibanda (2010). ""Just tell us what to do": Southern African face and its relevance to intercultural communication". In: *Journal of Pragmatics* 42.8, pp. 2158–2171.
Grice, Herbert Paul (1989). *Studies in the way of words*. Cambridge: Harvard University Press.

Groenendijk, Jeroen and Martin Stokhof (1991). "Dynamic predicate logic". In: *Linguistics and Philosophy* 14.1, pp. 39–100.
Haugh, Michael (2004). "Revisiting the conceptualisation of politeness in English and Japanese". In: *Multilingua* 23.1, pp. 85–109.
— (2007). "Emic conceptualisations of (im)politeness and face in Japanese: Implications for the discursive negotiation of second language learner identities". In: *Journal of Pragmatics* 39.4, pp. 657–680.
— (2010). "Jocular mockery, (dis)affiliation, and face". In: *Journal of Pragmatics* 42.8, pp. 2106–2119.
Horn, Laurence R (2004). "Implicature". In: *Handbook of Pragmatics*. Ed. by Laurence R Horn and Gregory Ward. Oxford: Blackwell, pp. 3–28.
Hymes, Dell (1962). "The ethnography of speaking". In: *Anthropology and Human Behavior*. Ed. by Thomas Gladwin and William C Sturtevant. Washington: Anthropology Society of Washington.
— (1963). "Introduction: Toward ethnographies of communication". In: *American Anthropologist* 66.6, pp. 1–34.
Kamp, Hans, Josef van Genabith, and Uwe Reyle (2011). "Discourse Representation Theory". In: *Handbook of Philosophical Logic*. Ed. by Dov M. Gabbay and Franz Guenthner. Vol. 15. Dordrecht: Springer Netherlands, pp. 125–394.
Kamp, Hans and Uwe Reyle (1993). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and Discourse Representation Theory*. Berlin: Springer.
Krifka, Manfred (2007). "Basic notions of information structure". In: *Interdisciplinary Studies of Information Structure*. Ed. by Caroline Féry, Gisbert Faneslow, and Manfred Krifka. Potsdam: Universität Potsdam, pp. 13–55.
Lewis, David (1969). *Convention: A philosophical study*. Harvard: Harvard University Press.
McCready, Eric (2015). *Reliability in Pragmatics*. Vol. 4. Oxford Studies in Semantics and Pragmatics. Oxford: Oxford University Press.
Mertens, Jean-François (1986). "Repeated games". In: *Proceedings of the International Congress of Mathematicians*. Ed. by Andrew M Gleason. Providence: American Mathematical Society, pp. 205–209.
Morrissey, Lochlan (2017). "On signalling games in improvised theatre". Griffith University.
Parikh, Prashant (1992). "A game-theoretic account of implicature". In: *TARK '92: Proceedings of the Fourth Conference on Theoretical Aspects of Reasoning About Knowledge*. San Fransisco: Morgan Kaufmann, pp. 85–94.
— (2001). "The use of language". In:
— (2006). "Pragmatics and games of partial information". In: *Game Theory and Pragmatics*. Ed. by Anton Benz, Gerhard Jäger, and Robert van Rooij. Basingstoke: Palgrave Macmillan, pp. 101–121.
— (2007). "Situations, rules, and conventional meaning: Some uses of games of partial information". In: *Journal of Pragmatics* 39.5, pp. 917–933.
Repp, Sophie (2012). "Common ground management: Modal particles, illocutionary negation, and VERUM". In: *Expressives and Beyond. Explorations of Conventional Non-truth-conditional Meaning*. Ed. by Daniel Gutzmann and Hans-Martin Gärtner. Oxford: Oxford University Press, pp. 231–274.
Rooij, Robert van (2003). "Conversational implicatures and communication theory". In: *Current and new directions in discourse and dialogue*. Ed. by Jan Kuppevelt and Robert W Smith. Berlin: Springer, pp. 283–303.

Rooij, Robert van (2008). "Games and quantity implicatures". In: *Journal of Economic Methodology* 15.3, pp. 261–274.
Russell, Stuart J and Peter Norvig (2010). *Artificial intelligence: A modern approach*. 3rd ed. Boston: Prentice Hall.
Silverstein, Michael (1993). "Metapragmatic discourse and metapragmatic function". In: *Reflexive language*. Ed. by John A Lucy. New York: Cambridge University Press, pp. 33–58.
— (2001). "The limits of awareness". In: *Linguistic anthropology: A reader*. Ed. by Alessandro Duranti. Malden: Wiley-Blackwell, pp. 382–401.
Soare, Robert I (2016). *Turing Computability*. Berlin: Springer.
Spencer-Oatey, Helen (2007). "Theories of identity and the analysis of face". In: *Journal of Pragmatics* 39.4, pp. 639–656.
Sperber, Dan and Deirdre Wilson (2004). "Relevance theory". In: *Handbook of Pragmatics*. Ed. by Laurence R Horn and Gregory Ward. Oxford: Blackwell, pp. 607–632.
Stalnaker, Robert C (1975). "Indicative conditionals". In: *Philosophia* 5.3, pp. 269–286.
— (2002). "Common ground". In: *Linguistics and Philosophy* 25.5, pp. 701–721.
Veltman, Frank (1996). "Defaults in update semantics". In: *Journal of Philosophical Logic* 25.3, pp. 221–261.
Weibull, Jörgen W (1997). *Evolutionary game theory*. Cambridge: MIT Press.