Mistakes About Conventions and Meanings

Mistakes About Conventions and Meanings

Cite this article as: Cosmo Grant, Mistakes About Conventions and Meanings, Topoi https://doi.org/10.1007/s11245-019-09656-3

Mistakes about conventions and meanings

Cosmo Grant

May 24, 2019

Abstract

The *Standard View* is that, other things equal, speakers' judgments about the meanings of sentences of their language are correct. After all, we make the meanings, so how wrong can we be about them? The Standard View underlies the *Elicitation Method*, a typical method in semantic fieldwork, according to which we should work out the truth-conditions of a sentence by eliciting speakers' judgments about its truth-value in different situations. I put pressure on the Standard View and therefore on the Elicitation Method: for quite straightforward reasons, speakers can be radically mistaken about meanings.

Lewis (1969) gave a theory of convention in a game-theoretic framework. He showed how conventions could arise from repeated *coordination games*, and, as a special case, how meanings could arise from repeated *signaling games*. I put pressure on the Standard View by building on Lewis's framework. I construct coordination games in which the players can be wrong about their conventions, and signaling games in which the players can be wrong about their messages' meanings. The key idea is straightforward: knowing your own strategy and payoff needn't determine what the others do, so leaves room for false beliefs about the convention and meanings. The examples are simple, explicit, new in kind, and based on an independently plausible meta-semantic story.

Title page

Author information

Paper title: Mistakes about conventions and meanings

Author: Cosmo Grant

Affiliation: Massachusetts Institute of Technology

Email: cosmodgrant@gmail.com

Address: 77 Massachusetts Avenue, Bldg 32-D808, Cambridge, MA, 02135, USA

Compliance with Ethical Standards

I declare that I have no conflict of interest.

This article does not contain any studies with human participants or animals performed by any of the authors.

1 Overview

The *Standard View* is that, other things equal, speakers' judgments about the meanings of sentences of their language are correct.[1] After all, we *make* the meanings, so how wrong can we be about them? I put pressure on the Standard View: for quite straightforward reasons, speakers can be radically mistaken about meanings.

Lewis (1969) gave a theory of convention in a game-theoretic framework. He showed how conventions could arise in repeated *coordination games*. He also introduced a special kind of coordination game, a *signaling game*, and showed how, as a special case of his theory, conventional meanings could arise in repeated signaling games.

I put pressure on the Standard View by building on Lewis's framework. I construct coordination games in which the players can be wrong about their own conventions. The key idea is simple: knowing your own strategy and payoff needn't determine what the others do, so leaves room for false beliefs about the convention. Guided by the coordination games, I consider the special case of signaling games, and construct signaling games in which the players can be wrong about their messages' meanings. We make the meanings, but we can still be wrong about them.

Perhaps we already know that speakers can be wrong about meanings, because of Twin Earth cases, or semantic deference, or the like ([Putnam, 1975], [Burge, 1979]). Still, the examples I give are interesting: they are simple, explicit, new in kind, and based on an independently plausible meta-semantic story.

Section 1 overviews the paper. Section 2 describes Lewis's game-theoretic framework. Sections 3–4 describe and discuss coordination games which leave room for mistakes about conventions. Section 5 shows that the Standard View is no straw man. Sections 6–7 describe and discuss signaling games which leave room for mistakes about meanings. Section 8 sums up.

2 Lewis's game-theoretic framework

2.1 A paradigm coordination game

Each day at noon, you and I play a game. You're the row-chooser: you play U or D. I'm the column-chooser: I play L or R. If we play UL or DR, we get lunch

[1]Thanks to Alex Byrne, Thomas Byrne, Nilanjan Das, Kevin Dorst, Justin Khoo, Vann McGee, Milo Phillips-Brown, Daniel Rothschild, Kieran Setiya, Robert Stalnaker, and two anonymous referees for helpful comments.

that day; else, we go hungry. Call this the *Simple Game*.

	L	R
U	1	0
D	0	1

Figure 1: Payoff matrix in the Simple Game

Imagine the first day we play the game. What will we do? No strategies suggest themselves. If I knew that you would play U, I would play L. If you knew that I would play R, you would play D. But neither of us knows how the other will play. More or less at random, you play U and I play R. We go hungry. Each thereby learns about the other. For example: you learn that I played R and I learn that you played U.[2]

The next day, we play again. What will we do? What we learned yesterday might help. If you think I'm stubborn, sticking with R, you will switch to D. If I think you're accommodating, switching to D, I will stick with R. But it might not help. If each of us is stubborn, or each thinks the other is accommodating, each of us will stick, and we'll go hungry again. If each of us is accommodating, or each thinks the other is stubborn, each of us will switch, and we'll go hungry again. Each of us uses what she learned yesterday, together with what she already knows about the other, to predict how the other will play, knowing that the other is doing the same.

Eventually, by good luck and good sense, we coordinate: you play U and I play L. The next day the memory of our success is fresh in our minds, and we repeat it. Success breeds success. We soon cease to worry. We coordinate day after day, each happy with her own choice and confident of the other's. We have established a convention: *play UL*.

2.2 A paradigm signaling game

Nature flips a coin and shows you the result. You then slip a note under my door, reading either # (pound) or & (amp). Then I have to guess how the coin landed. If I'm right, we get lunch that day; else, we go hungry. Each of us wants me to guess correctly. We must attempt this indirectly, by coordinating our strategies. Call this the *Coin Game*.

The first day, Nature flips the coin and it lands heads. You must either send me # or &, trying to signal how it landed so that I'll guess correctly. But which to send?

[2]Perhaps we learn more than this. You learn that I learn that you played U; I learn that you learn that I played R. And so on, up the hierarchy.

Imagine it. Neither message will be much help. So you send one at random, # say. Now I must guess. Your message was no help. So I guess at random, tails say. We go hungry. As before, we each thereby learn about the other. And as before, each of us uses what she learns, together with what she already knows about the other, to predict how the other will play, knowing that the other is doing the same.

Eventually, more or less by chance, we coordinate: the coin comes up heads, you send #, and I guess heads. We eat. When the coin comes up heads again the next day, our success is fresh in our minds, and we repeat it. When the coin comes up tails, what will we do? If we're sensible, surely you'll switch your message and then I'll switch my response. Success breeds success. Through good luck and good sense, we've fixed on complementary strategies: you send # when heads and & when tails; I guess heads given # and tails given &. We soon cease to worry. We coordinate day after day, each happy with her own choice and confident of the other's.

When we started playing the game we chose our moves more or less at random. In the course of playing the game over and over again, we have established a convention, and the messages have acquired meanings: # means *the coin landed heads* and & means *the coin landed tails*.

2.3 Taking Lewis's theory as a starting point

Lewis (1969) showed how conventions could arise from repeated coordination games, and, as a special case, how meanings could arise from repeated signaling games. I use Lewis's game-theoretic framework, and take his theory of convention as a starting point. But I don't endorse all the details of the theory. What I say about the examples in Section 3 conflicts with his theory, and the signaling games in Section 6 generalize his. I flag the differences as we go along. See Section 4.6 in particular. Here, I'll briefly discuss two worries about taking Lewis's theory as a starting point.

First worry. You might worry that small-scale interactions, like those in the Simple Game or the Coin Game, don't give rise to conventions. After all, if I make the coffee and my partner makes the eggs every morning, though we could equally well have done the reverse, it's unnatural to say that we have established a convention.[3]

Reply. First, I agree it's unnatural to say that my partner and I have established a convention. But I suggest that it's unnatural, not because of the scale (two people, low stakes), but because of other features, perhaps that, unlike in the Simple Game or the Coin Game, neither my partner nor I had to reason about

[3]Thanks to an anonymous referee for pressing this point and supplying the example.

the other in deciding what to do. Second, a range of theorists do think that even small-scale interactions, like those in the Simple Game and Coin Game, can be conventions. Take, for example, Hume's rowers (2000, 3.2.2) or Margaret Gilbert's walkers (1990b). Third, even if small-scale interactions don't give rise to conventions, that doesn't undermine the main claims of this paper. For the examples in the paper can easily be scaled-up, so that they involve lots of people and high-stakes situations. Fourth, I say with Lewis (p. 3) that "what I call convention is an important phenomenon under any name."[4]

Second worry. When we communicate, we do more than play signaling games—much more.

Reply. Meaning in signaling games is a simple example of linguistic meaning. A simple example is still an example; a special case is still a case. Signaling games are not *simplified models* of meaning; they are *simple cases* of meaning. (Compare: heredity in pea plants is a simple case of heredity, not a simplified model of heredity.) In studying signaling games, we have not changed the subject.

Here's Lewis:

> If we endow a hypothetical community with a great many [...] signaling conventions [where the messages are written or spoken] for use in various activities, with verbal expressions suitably chosen ad hoc, we shall be able to simulate a community of language users—say, ourselves—rather well. An observer who stayed in the background watching these people use conventional verbal messages as they went about their business might take a long time to realize that they were not ordinary language users. [...] Yet it remains true that our hypothetical verbal signalers do not do anything we do not do. We just do more. Their use of language duplicates a fragment of ours. (pp. 142–3.).

Lewis's hypothetical signalers don't do anything we don't do. We just do more. If the hypothetical signalers can be wrong about the meanings of their own messages, so can we. If they can be radically wrong, so can we.

3 Mistakes about conventions in coordination games

Conventions arise in repeated coordination games. As a special case, meanings arise in repeated signaling games. Before focusing on the special case of meanings, I consider conventions in general. This section shows how players can be

[4] All orphan page numbers refer to [Lewis, 1969]

wrong about their own conventions. I describe four games—the *Three-Player Game*, the *Five-Player Game*, the *Nature Game*, and the *Cycle Game*—and how each might turn out. I assume throughout that the structure of the situation is common knowledge, the players play repeatedly, and each player knows her own move and payoff but doesn't observe the other players' moves.

Each game leaves room for the players to be wrong about their own convention: on some ways things might turn out, the players establish a convention but are wrong about which. The key idea is simple: knowing your own move and payoff needn't determine what the others do, so leaves room for false beliefs about the convention.

The four examples are abstract. That helps make them clear and concise. But ordinary practical situations have the same structure as the examples. See Section 4.3. The examples are not mere theoretical curiosities.

3.1 First example: the Three-Player Game

Rowena, Colin and Mattea are playing a coordination game. Rowena is the row-chooser: she plays Up or Down (U or D). Colin is the column-chooser: he plays Left or Right (L or R). Mattea is the matrix-chooser: she plays West or East (W or E). If they play ULW or DRW or DLE or URE, each gets lunch; else, nothing.[5] Call this the *Three-Player Game*.

	L	*R*
U	1	0
D	0	1

W

	L	*R*
U	0	1
D	1	0

E

Figure 2: Payoff matrix in the Three-Player Game

Imagine the first day Rowena, Colin and Mattea play the game. What will they do? No strategies suggest themselves, and they play more or less at random: URW, say. They go hungry. Each thereby learns about the others. For example, Rowena learns that Colin and Mattea either played RW or LE.[6]

[5] I give the actions different labels: 'U' or 'D' for Rowena, 'L' or 'R' for Colin, 'W' or 'E' for Mattea. That makes things easier to follow. But different labels needn't mean different actions. For example, U, L, and W can all be the same action. See Section 4.3 for an example. Thanks to an anonymous referee for pressing me to clarify this.

[6] A warning. After the players make their moves, each gets a payoff—lunch, for example. The numbers in the matrices—*utilities*—represent the players' preferences over the payoffs. Utilities are coarser-grained than payoffs: if a player is indifferent between two payoffs (chicken and beef, say), then those payoffs have the same utility, even though the player may be able to tell apart the outcome in which they get the one (chicken) from the outcome in which they get the other (beef). Now, for the games in this paper, it matters which outcomes the players can tell apart. So the payoffs matter, not just the utilities. Therefore—and this is the key

The next day they play again. What will they do? What they learned yesterday might help. If Rowena knows that Colin and Mattea are stubborn, sticking with their strategies, then she will switch to D, so that they will coordinate next time. But it might not help. If Rowena doesn't know how Colin and Mattea will react, then she won't know how to react either.

They play the game day after day, learning about each other as they go. Eventually, by good luck and good sense, they coordinate: Rowena plays U, Colin plays L, Mattea plays W. The next day, the memory of their success is fresh in their minds, and they repeat it. Success breeds success. They coordinate day after day, each happy with her own choice and confident of the others'. They have established a convention: *play ULW*. So far, so familiar.

The structure of the game is common knowledge. Each player knows her move and payoff. But that's not enough for each player to work out what the others are doing. Take Rowena. She knows that she plays U and that she gets lunch every time. But for all she knows, Colin and Mattea could be playing LW or RE. And similarly for Colin and Mattea.

None of the players knows what the convention is. Still, they may have beliefs about it. Suppose that, for one reason or another, Rowena believes that the convention is *play URE*; Colin, that it's *play DLE*, Mattea, that it's *play DRW*. Section 4.1 explains why they might have these beliefs.

Take Rowena again. If you asked her what the convention is, she'd say, "it's *play URE*"; if she were to switch roles with Colin, she would play R and expect him to play U; she would bet at long odds that Colin and Mattea play RE.

As it happens, they're all wrong, for the convention is *play ULW*. Every player is wrong about the convention.

3.2 Second example: the Five-Player Game

A, B, C, D, and E are playing a coordination game. Each player has two actions: 0 or 1. The outcome 10101, for example, is the outcome in which A, C, E play 1 and B, D play 0. If they play 00000, 11000 or 11111, each gets positive payoff; else, nothing. Call this the *Five-Player Game.*

They play day after day. Eventually they coordinate on 00000, each happy with her own choice and confident of the others'. They have established a convention: *play 00000*.

point—in all the games interpret the utilities in the matrices as normal, as representing the players' preferences over the payoffs, except also assume that where the utilities are the same, the payoffs are the same too.

The structure of the game is common knowledge. Each player knows her own move and payoff. The only outcome consistent with A's move and payoff is 00000, and similarly for B. A and B know that the convention is *play 00000*. But C, D, E don't. For all they know, they could be playing 00000 or 11000.

C, D, E don't know what the convention is, but they may still have beliefs about it. Suppose that, for one reason or another, they believe, wrongly, that the convention is *play 11000*. Then a majority of players (C, D, E) have the same mistaken belief about their own convention.

3.3 Third example: the Nature Game

Roland and Col are playing a coordination game, but they're not sure which. Nature chooses the West Game or East Game at random. The game, once chosen, is fixed. Then Roland and Col play that game repeatedly. Roland chooses U or D; Col chooses L or R. Call this the *Nature Game*. The Nature Game is the two-player analog of the Three-Player Game, where Mattea's role has been taken by Nature, and Nature only gets to choose once.[7]

	L	*R*
U	1	0
D	0	1

West Game

	L	*R*
U	0	1
D	1	0

East Game

Figure 3: Payoff matrices in the Nature Game

As it happens, Nature chooses the West Game. Roland and Col play day after day. Eventually they coordinate on UL, each happy with her choice and confident of the other's. They have established a convention: *play UL*.

The structure of the situation is common knowledge.[8] Each player knows her move and payoff. As before, that's not enough for either to work out what the other is doing. So neither player knows the convention. But they may still have beliefs about it. Suppose that, for one reason or another, they believe they're in the East Game, so Ro believes the convention is *play UR* and Col believes it's *play DL*. Both are wrong.

[7] The Nature Game is known as a Bayesian game, because Roland and Col have incomplete information about the payoffs.

[8] The situation includes the initial choice by Nature. The game is either the West Game or East Game. The structure of the *game* is not common knowledge, since the players don't know which game Nature chooses. The structure of the *situation* is common knowledge.

3.4 Fourth example: the Cycle Game

Rosie and Colt are playing the Cycle Game.[9] Rosie chooses U, M or D; Colt chooses L, C or R.

	L	*C*	*R*
U	1	1	0
M	0	1	1
D	1	0	1

Figure 4: Payoff matrix in the Cycle Game

They play day after day. Eventually they coordinate on UL, each happy with her choice and confident of the other's. Have they established a convention? Unlike the previous examples, either player can unilaterally deviate from UL (Rosie, by playing D; Colt, by playing C) without losing out. Perhaps that means that *play UL* is not a convention. (Lewis thought so.) If so, strike out the example and jump to the next section. If not, I say that Rosie and Colt have established a convention: *play UL*.

As before, neither player knows the convention, but they may still have beliefs about it. Suppose that, for one reason or another, Rosie believes the convention is *play UC* and Colt believes that the convention is *play DL*. Both are wrong.

3.5 Further examples

The key idea behind all four examples is that knowing your own move and payoff needn't determine what the others do, so leaves room for false beliefs about the convention. The examples exploit that idea differently: in the Three-Player Game, every player is wrong, although wrong in different ways; in the Five-Player Game, a majority of players are wrong in the same way, although some are right; in the Nature Game, both players are wrong, because they are wrong about the payoff structure; in the Cycle Game, both players are wrong, assuming that the Cycle Game can give rise to conventions.

The idea applies generally. For example, can we find a game which leaves room for every player to be wrong about the convention *and* most of the players to have the same mistaken belief about the convention? Yes, it's not hard.[10] The recipe is simple. Pick your pattern of mistakes and cook up a game to match.

[9] Daniel Rothschild suggested this kind of example, but I don't mean to imply that he agrees with what I say about it.

[10] Suppose, for example, that there are eight players, each with two actions, 0 or 1. If an even number of players choose 1, all get lunch; else, nothing. Suppose they all eventually play 0, getting lunch every time, but, for one reason or another, two believe that they both play 0 and the other six play 1, while the six believe that those two play 1 and they all play 0.

4 Discussion of examples

4.1 Why the false beliefs?

In each example, some of the players don't know the others' choices. Take Rowena in the Three-Player Game: for all she knows, Colin and Mattea could be playing LW or RE. Even so, in the scenario I've described, she believes that they're playing RE. Why? Her belief seems to be unreasonable. And as with the Three-Player Game, so with the others.

Not so. On perfectly ordinary ways of fleshing out the scenarios, the players' beliefs about the conventions, although false, are reasonable. I'll focus on Rowena but what I say carries over to the other players and the other games.

4.1.1 Carelessness

Perhaps Rowena is careless, not realizing that her evidence is consistent with LW as well as RE. The more complicated the game, the more excusable the carelessness.

Careless people can establish conventions. If you and I reasoned sloppily in the Simple Game in coming to coordinate, that does not undermine our convention *play UL*. If Rowena, Colin and Mattea reason sloppily in the Three-Player Game, that does not undermine their convention *play ULW*.

4.1.2 False antecedent beliefs

Perhaps Rowena started the game with false beliefs about Colin and Mattea. For example: that Colin is stubborn and Mattea is accommodating; or that Colin wrongly thinks that Mattea is accommodating. Her antecedent beliefs, even if false, may be justified. Perhaps they're all friends. In any case, her antecedent beliefs, together with how things happen to go in the early rounds, lead her to believe that the convention is URE. Successive iterations reinforce her belief.

Or Rowena might have false beliefs, not about her opponents' *styles* of play (accommodating, or stubborn, or whatever), but about what *choices* they'll make. For example: that Colin will play R and Mattea will play E; or that Colin and Mattea will either play UR or LW.[11] Her antecedent beliefs may be

[11] The belief that Colin and Mattea will either play UR or LW won't lead Rowena to a false belief about the convention. I mention it just to point out that a player may have correlated beliefs about her opponents' actions. In other games, correlated beliefs may lead to false beliefs about the convention.

justified. Perhaps one outcome is salient. In any case, as before, they lead her to believe that the convention is URE.

Sometimes in game theory we assume that the players start from a position of radical uncertainty about each other. What exactly that assumption amounts to isn't always clear, but in any case it excludes the sort of antecedent beliefs I've been describing. However, the assumption is optional, not required. It's perfectly legitimate to suppose that the players have antecedent beliefs about each other. Kasparov didn't cease to play chess because he had true antecedent beliefs about what his opponent would do; I don't cease to play chess because I have false antecedent beliefs about what my opponent will do. We're not going beyond standard game theory by imagining that Rowena has the sort of antecedent beliefs I've been describing.

Even Rowena's being sure of what her opponents will do or having correlated beliefs about her opponent's choices, are perfectly consistent with standard game theory. In particular, such beliefs are perfectly consistent with the assumption that the players' choices are causally independent (and that causal independence is common knowledge).[12]

False antecedent beliefs about the others don't undermine a convention. If when we play the Simple Game you wrongly think I'm accommodating, or that I'll play R, that doesn't undermine our convention. Given Rowena's false antecedent beliefs, impeccable reasoning might lead her to believe that the convention is URE. Her antecedent beliefs don't undermine the convention. Impeccable reasoning doesn't undermine it either. So nor does the end result: her mistaken belief that the convention is URE.

4.2 Are the regularities conventions?

If the regularities in the examples are not conventions, then the examples don't show that people can be mistaken about their own conventions, for there are no conventions for them to be mistaken about. But the regularities in the examples are conventions. Again, I focus on the Three-Player Game but what I say carries over to the other examples.

[12] See [Stalnaker, 1998, pp. 43–4] for discussion of this point. To borrow one of his examples, suppose my partner and I are in our voting booths on election day. How she votes is causally independent of how I vote. You may have no idea how either of us will vote, but still be confident (and justifiably so) that, however we vote, we'll vote the same way.

4.2.1 First argument

If Rowena, Colin and Mattea established some convention or other in the Three-Player Game, then the regularity *play ULW* is a convention. They did establish some convention or other. Therefore the regularity *play ULW* is a convention.

To defend the first premise, suppose Rowena, Colin and Mattea established a convention other than ULW. What is it? A convention is, in particular, a regularity in behaviour. No outcome other than ULW is a regularity in behaviour. So no outcome other than ULW is a convention. Could the convention be some regularity other than an outcome of the game? I don't see what. If they established some convention or other, the regularity *play ULW* is a convention.

To defend the second premise, consider the similarities between the Simple Game and the Three-Player Game: the players' interests coincide; the players initially choose more or less at random; each uses what she learns, together with what she already knows about the others, to predict how the others will play in the knowledge that they are doing the same; several outcomes yield the preferred payoff; eventually, the players settle on actions, each happy with her own and confident of the others'; if anyone deviates, nobody benefits.

The Simple Game leads to a paradigm convention. The Three-Player Game resembles the Simple Game both in structure and game-play. The Three-Player Game leads to a convention too.

4.2.2 Second argument

Here's a happier way the Three-Player Game might turn out. Each player believes that the convention is ULW. Their beliefs are justified, for they started the game with justified antecedent beliefs about the others. Contrast that with how things actually turn out. Rowena wrongly believes the convention is URE; Colin, that it's DLE; Mattea, that it's DRW. Perhaps they were careless; or perhaps they had false antecedent beliefs about the others.

In the happier case, each player believes that *play ULW* is the convention. They're correct, and not by luck. Each started the game with true beliefs about the others (say, that Colin is stubborn and Mattea is accommodating). Their true antecedent beliefs, together with how things happen to go in the early rounds, lead them to believe that *play ULW* is the convention. Perhaps they're not in a position to know that *play ULW* is the convention. Nevertheless, their beliefs are reasonable. It's not a requirement on conventions that you only believe what you're in a position to know. That would be absurdly demanding. The regularity *play ULW* is a convention in the happier case. Don't punish the players for reasoning well about each other.

Back to the actual case: each player is mistaken about the convention, because of carelessness or false antecedent beliefs or whatever. Conventions don't depend on how attentive the players are, nor on how well they know each other. If the regularity *play ULW* is a convention in the happier case, it's a convention in the actual case too.

Putting things together, I conclude that the regularity *play ULW* is a convention in the actual case, as required.

4.3 Concrete examples

The four examples are abstract. That makes them clean, clear and concise. They may also seem contrived. But they aren't. For ordinary practical situations have the same structure as the abstract examples and could easily turn out in the same way. To dismiss the examples as mere theoretical curiosities is a mistake.[13]

4.3.1 Concrete example for the Three-Player Game

Take the Three-Player Game. Suppose Rowena, Colin and Mattea all work in the same restaurant kitchen. At about 6pm each day, the chef puts two trays of squash in the oven to roast, one on top and one on bottom. After twenty minutes or so the trays need to be swapped, else the top one will burn and the bottom one won't caramelize. Swapping the trays is up to Rowena, Colin or Mattea, but it's one task among hundreds in a hectic kitchen and it isn't clear which of them will do it.

Imagine it. If none of them swaps the trays, the squash will be ruined. If just one of them does, it'll be fine. More is true. If two of them swap the trays in turn, neither realizing the other's plan, the trays will end up back where they started and the squash will be ruined. Similarly, if all three of them swap the trays in turn, it'll be fine. Of course, if all three of them swap the trays some effort is wasted, but that's small fry compared to roast squash.

The situation has the same structure as the Three-Player Game. U and D correspond to Rowena's swapping and not swapping the trays, and similarly for L and R, and W and E. In four outcomes (URE, DLE, DRW, where only Rowena or Colin or Mattea swap the trays, and ULW, where all three do), the squash is fine. In the other four outcomes (DRE, where none of them does, and URW, DLW, ULE, where two of them do), the squash is ruined.

What might happen? Here's one way things might go. At first, each of Rowena, Colin and Mattea thinks that one of the others will swap the trays. So no one

[13] Thanks to an anonymous referee for pressing me on this point.

does, and the squash is ruined. When the chef shouts at them they realize what happened and in the heat of the kitchen each goes away thinking that she alone will swap the trays from now on. The next day all three swap them, and the squash is fine. The kitchen is hectic and none of them sees the others do it. Their mistaken beliefs are reinforced. And so it goes on, each happily swapping the trays and confident that she alone is doing so. They have established a convention, *all swap the trays*, but each is wrong about what the convention is.

4.3.2 Concrete example for the Nature Game

Or take the Nature Game. Suppose Roland and Col are each trying to schedule a departmental reading group on Mondays. Roland's can start at 9am or 2pm; Col's at 11am or 4pm. Reading groups last two hours. Roland and Col are rivals and, out of pride, won't attend each other's group nor even coordinate times directly.

If the groups meet at 9am and 4pm, or 11am and 2pm, there'll be a break in between. If they meet at 9am and 11am, or 2pm and 4pm, there won't be. Roland and Col aren't sure which is best. On the one hand, four hours is a long time to concentrate. So if there is a break, maybe attendance will be higher, because most people will attend both. On the other hand, it's useful to have an uninterrupted morning or afternoon. So if there's a break, maybe attendance will be lower, because few people will attend both.

The situation has the same structure as the Nature Game. U and D correspond to 9am and 2pm; L and R correspond to 11am and 4pm. If a break is better, Roland and Col are in the East Game; if no break is better, they're in the West Game. Whether or not a break is better depends on people's preferences, which Roland and Col aren't sure about.

What might happen? Here's one way things might go. As it happens, no break is better. The first week, Roland's group meets at 9am and Col's at 4pm. Attendance is low. Roland suspects, wrongly but reasonably, that a break is better and that Col's group met at 11am. But, being stubborn, he sticks with 9am. Col suspects, wrongly but reasonably, that a break is better and that Roland scheduled his group for 2pm. And, being pragmatic, she switches to 11am. The next week, attendance is high. Roland and Col's mistaken beliefs are reinforced. And so it goes on, each happy with her own group's time and confident of the other's. They have established a convention, *meet at 9am and 11am*, but each is wrong about what the convention is.

4.4 Belief about regularities and belief about conventions

Distinguish two claims: Rowena believes that the regularity is *play URE*; Rowena believes that the convention is *play URE*. I've glossed over the difference. But in fact the claims are independent. Rowena might believe that the convention but not the regularity is *play URE*, because she might mistakenly believe that a convention doesn't require a regularity. She might believe that the regularity but not the convention is *play URE*, because she might mistakenly believe that the structure of the Three-Player Game rules out conventions.

I assume that whenever a player believes the regularity is *play such-and-such*, the player believes the convention is *play such-and-such*. The assumption isn't true in general. All I claim is that the four examples might turn out that way. The examples leave room for mistakes about conventions. They don't force mistakes.

4.5 Mistakes about conventions on the cheap

Suppose again that Rowena believes wrongly that the structure of the Three-Player Game rules out conventions. Perhaps she's a philosopher in the grip of a false theory of convention. And suppose as before that each player, including Rowena, does her bit of URE, happy with own choice and confident of the others'. Then the regularity *play URE* is a convention but Rowena believes it isn't. She's mistaken about her own convention.

Or suppose that each day when it's time to make her move Rowena plays U but afterwards forgets her choice, believing she played D. Perhaps conventions can survive this selective forgetfulness. Except when she makes her move, Rowena believes the convention is, say, *play DLE*; in fact, it's *play ULW*. Most of the time she's mistaken about her own convention.

Or suppose Rowena lacks the concept of a convention, so although she correctly believes that the regularity is *play ULW*, she doesn't believe that it's a convention.

These are mistakes about conventions on the cheap, relying on fussy details or exotic situations. The examples in Section 3 are abstract in order to make the structure clear. They are not fussy; they are not exotic.

4.6 How do the examples fit with Lewis's theory?

Are my claims about conventions sanctioned by Lewis's theory? No. This section spells out the details of his theory and shows why, according to it, my examples are not examples of conventions. So Lewis's theory is wrong.

We need some preliminary definitions. A *strategy profile* is a tuple of strategies, one for each player. A strategy profile is a *Nash equilibrium* if, for each agent, if she alone had done otherwise, she would be no better off. A strategy profile is a *coordination equilibrium* if, for each agent, if she alone had done otherwise, no one would be better off. In a Nash equilibrium, no one wishes that she alone had done otherwise ('no regret'); in a coordination equilibrium, no one wishes, of any one else, that she alone had done otherwise ('no resentment'). A strategy profile is a *proper coordination equilibrium* just if, for each agent, if she alone had done otherwise, no one would be better off *and someone would be worse off*.[14] A *coordination problem* is a situation of interdependent decision by two or more agents in which their interests largely coincide and which has two or more proper coordination equilibria.

Here is Lewis's first pass at a theory of convention.[15] A regularity R in agents' behaviour when in a recurrent situation S is a *convention* if and only if, in any instance of S,

(1) everyone conforms to R;

(2) everyone expects everyone else to conform to R;

(3) everyone prefers to conform to R on condition that the others do, since S is a coordination problem and uniform conformity to R is a proper coordination equilibrium in S.

How do my examples fit with Lewis's theory? Well, all four are situations of interdependent decision by two or more agents in which interests coincide. In the Cycle Game, there are three coordination equilibria, but none is proper; and the definition of coordination equilibrium doesn't apply to the Nature Game. So the regularities in these games are not conventions, on Lewis's theory.

In the Three- and Five-Player Games, there are two or more proper coordination equilibria. But the regularities aren't conventions, according to Lewis's theory, for another reason.

Take the Three-Player Game. Everyone conforms to the regularity *play ULW*. So (1) is true. And everyone prefers to conform to that regularity on condition that the others do. So (3) is true. But not everyone expects everyone else to conform to it. For example, Rowena expects Colin and Mattea to do RE, not LW. (Rowena does expect the others to conform to what she takes to be the actual regularity, URE, but she doesn't expect the others to conform to what is in fact the actual regularity, ULW.) So (2) is false. Therefore the regularity *play ULW* is not a convention, on Lewis's theory.

[14] Gilbert (1981) pointed out that the term 'proper coordination equilibrium' is ambiguous, and Lewis didn't make clear which he intended. Gilbert reports that Lewis clarified in private communication that this is what he had in mind.

[15] p. 42

In short: according to Lewis's first pass at a theory of convention, none of the regularities in my four examples is a convention. Lewis's final theory is more complicated.[16] But the complications don't change things: on his final theory, too, the regularities are not conventions.

If Lewis's theory is correct, the examples don't show you can be wrong about your own conventions, for the examples are not examples of conventions. But Lewis's theory isn't correct. As I've argued, the regularities in the examples are conventions. Lewis provided a clear and simple framework, and he brought to light significant features of conventions. But not all the details of his theory are correct.

I don't have a replacement in mind. One could look for a minimal departure from Lewis's theory according to which the examples are examples of conventions. But that's not a profitable line of inquiry, since Lewis's theory is questionable in other respects too, like his insisting on a proper coordination equilibrium. (See [Gilbert, 1981], [Gilbert, 1983] and [Vanderschraaf, 1998] for prominent criticisms.) Developing a theory of convention would support my argument, but it isn't essential.

5 Motivating The Standard View

In Sections 3–4, I described coordination games which, for quite straightforward reasons, leave room for mistakes about conventions. In Sections 6–7, I consider the special case of signaling games, and describe signaling games which leave room for mistakes about meanings. These games put pressure on the *Standard View* that, other things equal, speakers' judgments about the meanings of sentences of their language are correct. This section motivates the next two by showing that the Standard View is not a straw man.

5.1 Language mavens

Steven Pinker devotes a chapter of his book *The Language Instinct* to criticizing language mavens, those self-appointed authorities on usage, who pull people up for using 'who' instead of 'whom', saying 'very unique', confusing 'disinterested' and 'uninterested', and the like. Pinker is concerned with syntax, not semantics, but the issues are parallel. Here's how the chapter starts:

> Imagine that you are watching a nature documentary. The video shows the usual gorgeous footage of animals in their natural habitats. But the voiceover reports some troubling facts. Dolphins do

[16] p. 78

> not execute their swimming strokes properly. White-crowned sparrows carelessly debase their calls. Chickadees' nests are incorrectly constructed, pandas hold bamboo in the wrong paw, the song of the humpback whale contains several well-known errors, and monkeys' cries have been in a state of chaos and degeneration for hundreds of years. Your reaction would probably be, What on earth could it mean for the song of the humpback whale to contain an "error"? Isn't the song of the humpback whale whatever the humpback whale decides to sing?

He continues:

> To a linguist or psycholinguist, of course, language is like the song of the humpback whale. The way to determine whether a construction is "grammatical" is to find people who speak the language and ask them. [Pinker, 1995, p. 370]

As with syntax, so with semantics: to determine whether a construction is grammatical, find people who speak the language and ask them; to determine what a sentence means, find people who speak the language and ask them.

To find out the meaning of 'literally', or 'decimate', or 'enormity', or the other favourites of the language mavens, don't argue from the armchair, nor fixate on etymology—just ask people! Only a language maven would disregard the judgments of ordinary speakers. After all, we make the meanings, so how wrong can we be about them? Other things equal, implies Pinker, speakers' judgments about meanings are correct.

5.2 The Elicitation Method

How should we work out the truth-conditions of a sentence? Here is the *Elicitation Method*: Describe scenarios and ask many speakers whether the sentence is true relative to each scenario. If the speakers judge that the sentence is true relative to a scenario, the sentence *is* true relative to that scenario, or in other words the scenario does belong to the sentence's truth-conditions. If the speakers judge that the sentence isn't true relative to a scenario, the sentence *isn't* true relative to that scenario, or in other words the scenario doesn't belong to the sentence's truth-conditions.

We could refine the Elicitation Method in all sorts of ways: ask only native speakers; instead of asking the speakers for a binary judgment (whether the sentence is true relative to a scenario), ask them for a graded judgment (how well the sentence fits a scenario); instead of asking about one sentence, ask about lots of sentences of the same form; instead of describing the scenarios in the same language as the sentence, use another language, or use pictures or videos; help

the speakers distinguish infelicity from falsehood; avoid asking speakers who are corrupted by theory (semanticists, for example); add filler questions so as to obscure the experiment's purpose; randomize the order of the questions...

The Elicitation Method, or some refinement of it, is a standard method in semantic fieldwork. Take Altshuler et. al. (2019, Chapter 1): "We will judge our progress in terms of how closely the system we develop tracks the intuitions speakers have about the truth of a sentence in different situations." Or Winter (2016, p. 16): "Just as intuitive judgments about sentence grammaticality have become a cornerstone in syntactic theory, intuitions about entailments between sentences are central for natural language semantics." (Winter uses judgments about entailments, not about truth-conditions, but the approaches are equivalent.) Or Matthewson (2004, p. 369): "direct elicitation (including asking consultants for judgments) is an indispensable methodological tool." For thorough discussion, see [Matthewson, 2004] or [Bochnak and Matthewson, 2015].

The Elicitation Method is not the only way to work out the truth-conditions of a sentence. Other techniques are available. For example, you might gather texts or record conversations and extract truth-conditions from patterns of use. Or you might ask bilingual speakers to translate a sentence of the language under study into another language. And so on. Still, there is no question that eliciting speakers' judgments is a standard method in semantic fieldwork. The Standard View justifies the Elicitation Method. By better understanding how speakers' judgments about meanings can go wrong, we will better understand the limits of the Elicitation Method.

5.3 Lewis on knowledge of conventions

According to Lewis, participants in a convention are in a position to know what the convention is. As with conventions in general, so with conventions of language in particular: speakers of a language are in a position to know what their linguistic conventions are, or in other words, to know the meanings of their terms.

Lewis tempers the claim by pointing out that you may be in a position to know the convention without actually knowing it, that you may not be able to put what you know into words, and that snap judgments about the convention, like snap judgments about anything, may be wrong.[17] Still, other things equal, speakers' judgments about meanings are correct.

A view held by Lewis is a view worth taking seriously. In Sections 3–4, I argued that Lewis's view is wrong: participants in a convention may fail to know, or even be in a position to know, what the convention is; they may believe of some other

[17] He describes a further qualification, too, involving Abelard's distinction between beliefs and expectations *in sensu composito* and *in sensu diviso*. See pp. 64–8.

regularity that it's the convention; they may easily state their mistaken belief verbally; they may stand by their mistaken belief even after careful reflection. In Sections 6–7, I argue that Lewis's view is wrong about conventions of language in particular.

6 From coordination games to signaling games

It's easy to come up with coordination games which leave room for mistakes about conventions. Remember the key idea: knowing your own move and payoff needn't determine what the others do, so leaves room for false beliefs about the convention.

A signaling game is a special kind of coordination game. Signaling games are particularly interesting, since they give rise to meanings. In this section I apply the key idea to the special case of signaling games. I construct signaling games which leave room for mistakes about meanings.

Just as the examples from Section 3 go beyond the Simple Game, our paradigm coordination game, so too the examples in this section go beyond the Coin Game, our paradigm signaling game. We must be careful, when we go beyond the Coin Game, to ensure that a Lewis-style analysis of the messages' meanings still applies. Our examples must balance two desiderata: on some ways the game can turn out, the messages have meanings; and the players can be mistaken about the meanings. The first desideratum pulls us towards the Coin Game, for a Lewis-style analysis of meanings is most straightforward in games like that. The second desideratum pushes us away from the Coin Game, for games like that leave little room for mistakes about meanings. See Section 7.1 for further discussion.

I describe four games—the *ABC Game*, the *Two-Sender Coin Game*, the *Coin Game with Nature*, the *Signaling Cycle Game*—and how each might turn out. I assume as in Section 3 that the structure of the situation is common knowledge, the players play repeatedly, and each player knows her own move and payoff but doesn't observe the other players' moves. Each game leaves room for the players to be wrong about the messages' meanings: on some ways things might turn out, the messages acquire meanings but some players are wrong about some meanings.

Remember Lewis's hypothetical verbal signalers. They don't do anything we don't do. We just do more. If the hypothetical signalers can be wrong about the meanings of their own terms, so can we. If they can be radically wrong, so can we.

State	Sienna	Reg	Rae	Roy	Payoff
A	!	A	C	B	1
B	&	C	B	A	1
C	#	B	A	C	1

Table 1: Actual strategies in the ABC Game.

6.1 The ABC Game

Sienna, Reg, Rae, and Roy are playing a signaling game. Sienna is the sender; Reg, Rae and Roy are receivers. Nature chooses one of three states—A, B, C—at random. Sienna observes the state. She sends one of three messages, ! (bang), & (amp), # (pound), to the receivers (the same message to each). Then the receivers independently guess the state. If at least one guesses correctly, they all get positive payoff; else, nothing.

They play the game day after day. Eventually, they coordinate. Their strategies are represented in Table 1. For example: Sienna sends ! when A, & when B, # when C; Reg guesses A given !, C given &, B given #. Each receiver guesses correctly given one of the messages (Reg given !, Rae given &, Roy given #) but incorrectly given the other two. Since someone guesses correctly no matter what the state, everyone always gets positive payoff. Each is happy with her own strategy and confident of the others'. The messages have acquired meanings: ! means A, & means B, # means C.

No receiver knows what Sienna is doing. Still, they may have beliefs about it, and corresponding beliefs about the messages' meanings. Suppose each is cocky, believing he always guesses correctly. For example, Reg believes Sienna sends ! when A, # when B, & when C, and so believes that ! means A, # means C, & means B. And similarly for Rae and Roy. (See Section 7.2 for further discussion.)

Sienna knows the meanings: she knows her own strategy and the meanings are determined by that. (See Section 7.3 for further discussion.) The receivers don't. Reg is right about the meaning of ! and wrong about the meanings of & and #. And similarly for Rae and Roy. All receivers are wrong about the meanings of two messages. For each message, a majority of receivers (two of three) are wrong about the message's meaning.

Since there are as many receivers as states, if the receivers guess differently, one of them is bound to guess correctly. If any receiver guesses correctly, all get the preferred payoff. So, if the receivers could confer among themselves, they could make sure to guess differently given each message, and so ensure they get the preferred payoff every time, regardless of Sender's strategy. If that were what happened, perhaps the messages wouldn't be meaningful. But that's not what happens. The receivers don't confer among themselves. The fact (if it is a

fact) that if they were to confer the messages wouldn't be meaningful doesn't undermine the claim that the message are meaningful, given that the receivers don't confer.

6.2 The Two-Sender Coin Game

Nature flips two coins. Sender 1 sees how the first coin landed and sends a message, m_1 or m_2, to Receiver; Sender 2 sees how the second coin landed and sends a message, m_3 or m_4, to Receiver. The senders act independently. After receiving the messages, Receiver guesses how each coin landed. If she gets both right, everyone gets positive payoff; else, nothing.

They play day after day. Eventually, they coordinate: Sender 1 sends m_1 when heads, m_2 when tails; Sender 2 sends m_3 when heads, m_4 when tails; and Receiver's strategy complements theirs, so she always guesses correctly. The messages have acquired meanings: m_1 and m_3 mean *the coin landed heads*, m_2 and m_4 mean *the coin landed tails*.[18]

The structure of the game is common knowledge. Each player knows her strategy and payoffs. That's enough for Receiver to work out the senders' strategies. And it's enough for each sender to work out how Receiver responds to his messages. But it's not enough for either sender to work out the other sender's strategy, nor how Receiver responds to the other sender's messages. For all Sender 1 knows, Sender 2 might send m_4 when heads and m_3 when tails, and Receiver guess heads given m_4 and tails given m_3. And similarly for Sender 2.

Still, each sender may have beliefs about the other sender's strategy, and corresponding beliefs about the messages' meanings. Suppose each sender flips the other's strategy, so Sender 1 believes that m_3 means tails and m_4 means heads, and Sender believes that m_1 means tails and m_2 means heads. Each is wrong about the meanings of the other's messages.

6.3 The Coin Game with Nature

Sender and Receiver are playing a signaling game, but they're not sure which. Nature chooses the West Game or the East Game at random. The game, once chosen, is fixed. Then Sender and Receiver play that signaling game repeatedly.

In either game, Nature flips a coin and Sender sees the result. Sender sends a message, # or &, to Receiver, who then guesses how the coin landed. In the

[18] Or perhaps m_1 means *the first coin landed heads* and m_2 means *the second coin landed heads*, and so on. We needn't decide the matter here.

East Game, the players are rewarded if Receiver guesses correctly; in the West Game, the players are rewarded if Receiver guesses incorrectly.[19]

	guess H	guess T
H	0	1
T	1	0

West Game

	guess H	guess T
H	1	0
T	0	1

East Game

Figure 5: State-response correspondences in the Coin Game with Nature

As it happens, Nature chooses the East Game. Sender and Receiver play day after day. Eventually they coordinate: Sender sends # when heads and & when tails; Receiver guesses heads given # and tails given &. Each is happy with her strategy and confident of the other's. The messages have acquired meanings: # means *the coin landed heads* and & means *the coin landed tails*.

As before, neither player knows the other's strategy. Still, they may have beliefs about it, and corresponding beliefs about the meanings. Suppose they believe Nature chose the West Game, so Sender believes Receiver guesses tails given # and heads given & and Receiver believes Sender sends # when tails and & when heads. In short: each player is wrong about the game and flips the other's strategy.

Sender, despite his mistake, knows that # means heads and & means tails, since he knows his own strategy. Receiver is not so lucky: she believes that # means tails and & means heads.

6.4 The Signaling Cycle Game

Nature chooses one of three states (s_1, s_2, or s_3) at random. Sender observes the state and sends a message (m_1, m_2, or m_3) to Receiver. Then Receiver chooses a response (r_1, r_2, or r_3). The payoffs for each state-response pair are given below:

	r_1	r_2	r_3
s_1	1	1	0
s_2	0	1	1
s_3	1	0	1

Figure 6: State-response correspondence in the Signaling Cycle Game

Sender and Receiver play day after day. Eventually they coordinate: Sender sends m_2 when s_1, m_3 when s_2, m_1 when s_3; Receiver does r_1 given m_1, r_2

[19]Note that the matrices don't represent the games in strategic form; rather, they represent the payoffs for each state-act pair, from which the strategic forms may be derived, given a probability distribution over the states.

given m_2, r_3 given m_3. Each is happy with her strategy and confident of the other's.

Have the messages acquired meanings? Unlike the previous examples, for each state there are two responses which yield the preferred payoff. Therefore each player could unilaterally change her strategy without losing out. Perhaps that means the messages don't acquire meanings. If so, strike out the example and jump to the next section. If not, I say that m_1 means s_3, m_2 means s_1, m_3 means s_2.

As above, neither player knows what the other is doing, but they may have beliefs about it, and corresponding beliefs about the meanings. Suppose each player permutes the other's strategy. Sender, despite his mistake, knows that m_1 means s_3, m_2 means s_1, m_3 means s_2, since he knows his own strategy. Receiver is not so lucky: she believes that m_1 means s_1, m_2 means s_2, m_3 means s_3. She is wrong about the meaning of every message.

7 Discussion of examples

7.1 Why go beyond basic signaling games?

In a *basic signaling game*, like the Coin Game, there are two players: Sender and Receiver. One of n possible states, $s_1, \ldots, s_n$, is chosen by Nature with equal probability. Sender observes the state; Receiver doesn't. Sender sends one of n possible messages, $m_1, \ldots, m_n$, to Receiver, who then chooses one of n responses, $r_1, \ldots, r_n$. The payoffs depend on the state and the response. If Receiver does r_i in s_i, Sender and Receiver each get equal positive payoff. Otherwise, each gets nothing.[20] A strategy for Sender is a function from states to messages. A strategy for Receiver is a function from messages to responses. Both players want Receiver to guess correctly. They must attempt this indirectly, by coordinating their strategies.

The four examples in Section 6 go beyond basic signaling games in all sorts of ways: in the ABC Game, there is one sender and several receivers, only one of whom has to guess right; in the Two-Sender Coin Game, there are two senders and one receiver, who has to guess how both coins landed; in the Coin Game with Nature, the players are unsure about the state-response correspondence; the Signaling Cycle Game relaxes the payoff structure.

Why go beyond basic signaling games? Because basic signaling games leave little

[20] Isn't it obvious what the players should do, namely, send m_i in s_i and do r_i given s_i? No. That confuses a property of our representation (how we label the states, messages and responses) with a property of what we're representing.

room for mistakes about meanings. Take the Coin Game, a basic signaling game with two states, messages and responses. The messages acquired meanings. You know your strategy and payoffs. That is enough to work out my strategy. I know my strategy and payoffs. That is enough to work out your strategy. Since each of us can work out the other's strategy, each of us can work out what the messages mean. The Coin Game doesn't leave room for mistakes about meanings.

Now consider a larger basic signaling game, say with ten states, messages and responses. Here's one way things might turn out. Sender and Receiver play day after day, learning about each other as they go. They eventually settle on strategies, each happy with her own and confident of the other's. In states $s_1, \ldots, s_8$ their strategies match up: Sender sends m_i in s_i and Receiver does r_i given m_i. These messages have acquired meanings: m_i means s_i ($i = 1, \ldots, 8$). In states s_9 and s_{10}, their strategies don't match up: Sender sends m_9 in s_9 and m_{10} in s_{10}, but Receiver does r_{10} given m_9 and r_9 given m_{10}. By good luck, Nature doesn't choose s_9 or s_{10}. The players believe, wrongly but with good reason, that their strategies would match up no matter the state.

It's not clear whether m_9 and m_{10} are meaningful. If they are meaningful, then m_9 means s_9 and m_{10} means s_{10} (although see Section 7.5). Whether meaningful or not, we may suppose that Receiver believes, wrongly, that m_9 means s_{10} and m_{10} means s_9. Basic signaling games do, thus, leave room for mistakes about meanings.

I don't rely on examples like this. The example depends on an unlikely event (that Nature doesn't choose two of the states). More importantly, the players' beliefs are not robust: with probability 1, eventually Nature will choose s_9 or s_{10}, and then the players will correct their mistakes. If they're mistaken about meanings, they won't be for long. The examples in Section 6, by contrast, don't depend on unlikely events, nor need the players ever realize their mistakes.

When conditions are strange enough (the players are fantastically unlucky, or selectively forgetful, or dazed and confused, or philosophers), no doubt they can be mistaken about meanings. Examples like that aren't interesting. The aim is not just to find signaling games which leave room for mistakes about meanings, but to find games which leave room for mistakes about meanings in a simple, straightforward way. You don't need outlandish set-ups to be mistaken about meanings. The examples in Section 6 are artificial, in order to make the structure clear. They are not fussy; they are not exotic.

7.2 Belief about regularities and belief about meanings

Consider, say, Reg in the ABC Game. Distinguish two claims: (a) Reg believes Sienna sends ! when A, # when B, & when C; (b) Reg believes that ! means A, # means B, & means C. I've glossed over the difference. But in fact the

claims are independent: you might have (a) without (b), if Reg (mistakenly) believes that messages in a signaling game can't acquire meanings; you might have (b) without (a), if Reg (mistakenly) believes that meanings don't require a regularity.

I've assumed that by the time the players coordinate, each happy with her own strategy and confident of the others', if a player believes that a sender sends message m in state s, she also believes that m means s. The assumption isn't true in general. All I claim is that the four examples might turn out that way. The examples leave room for mistakes about meanings; they don't force mistakes.

7.3 Sender is not wrong about meanings

The meaning of a message, given that it's meaningful, is determined by Sender's strategy. Look back at the examples: the meaning of a message, given that it's meaningful, is the state in which Sender sends that message.

Suppose m means s. Since m is meaningful, the players coordinate, each happy with her own strategy and confident of the others'. Since m means s, Sender sends m whenever the state is s. Sender knows his strategy. Hence Sender knows that he sends m in s. Given the assumption stated in the previous subsection, it follows that Sender believes that m means s. Generalizing, Sender is right about the meanings of his own messages, if they're meaningful at all. The key idea—that knowing your own strategy and payoff needn't determine what the others do—doesn't leave room for Sender to be wrong about the meanings of his own messages.

When there are multiple Senders, each Sender may be wrong about the meanings of the *other* Senders' messages, as in the Two-Sender Coin Game. But each Sender still knows the meanings of his own messages. That is a limitation of the examples.

7.4 Switching roles

In all the examples, the players' roles are fixed. In particular, senders never become receivers and receivers never become senders. In actual languages, of course, things aren't like that. People sometimes speak and sometimes listen. No problems arise in actual languages from switching roles. The conventions of meaning in actual languages are robust to role-switches.[21]

Is the same true of the conventions of meaning in the examples? The short answer is: typically not. In each case I assume, naturally enough, that a player

[21] Thanks to an anonymous referee for pressing this point.

behaves in a new role as she believes the player she is taking over from behaved. Now, take the Signaling Cycle Game. Suppose Sender and Receiver swap roles. Then Receiver, now in Sender's role, will send m_1 when s_1, m_2 when s_2 and m_3 when s_3. And Sender, now in Receiver's role, will do r_3 given m_1, r_1 given m_2 and r_2 given m_3. The result? They'll get zero payoff no matter the state. Swapping roles leads them to anti-coordinate. The situation is similar for the Two-Sender Coin Game and the ABC Game.[22] As it happens, in the Coin Game with Nature, the players can switch roles without any problem arising. But typically the conventions of meaning in the examples are not robust to role-switches. That is another limitation of the examples.

Still, it's worth noting that the coordination game examples from Section 3 are more robust to role-switches. Take the Three-Player Game. If any *two* players switch roles (Colin becoming the matrix-chooser and Mattea becoming the column-chooser, say) then no problems will arise. But if all *three* players switch roles, they'll get zero payoff. Similarly, the Five-Player Game is robust to many, but not all, role-switches. The Nature Game is robust to the two players' switching roles. The Cycle Game isn't.

The conventions in the coordination games are often, but not always, robust to role-switches, even when the players are wrong about the convention. That is evidence—even if only weak evidence—that conventions of meaning can be robust to role-switches too, even when the players are wrong about the meanings.

7.5 Indicative meanings and imperative meanings

Recall the Coin Game: you send # when heads and & when tails; I guess heads given # and tails given &. We coordinate day after day, each happy with her own strategy and confident of the other's. I said that # means *the coin landed heads* and & means *the coin landed tails*. These are the *indicative* meanings of the messages.

As Lewis pointed out, there's an alternative interpretation: # means *guess heads* and & means *guess tails*. These are *imperative* meanings of the messages. An indicative meaning gives information about the state. An imperative meaning gives an instruction about the response.

You might take the imperative meaning of a message to be determined by Sender's strategy, just like the indicative meaning: the imperative meaning of m is to make the appropriate response, whichever it is, to the state in which Sender sends m. Or you might take the imperative meaning to be determined by Receiver's strategy: the imperative meaning of m is to respond however Re-

[22] For the ABC Game, things are not completely straightforward, because we don't know what Sienna will do as a receiver. But if we suppose that she behaves the same way no matter which receiver she swaps with, then things will go wrong.

ceiver does respond given m. In basic signaling games, these coincide; in more complicated games, they may not. Taking the meanings to be indicative, Sender is in a privileged position; taking the meanings to be imperative, Receiver is in a privileged position.

When should we interpret messages as indicative and when as imperative? It's not clear. In these simple settings, either interpretation may be acceptable.[23] If we interpret the messages one way, we may get one pattern of mistakes about the meanings; if we interpret them the other way, we may get another.

The distinction between indicative and imperative meanings might help leave room for Sender to be wrong about the meanings of his own messages, despite knowing his own strategy, for he could be wrong about the *imperative* meanings, as determined by *Receiver's* strategy. But better to find examples which don't rely on choosing between indicative and imperative meanings. And better to find meanings which every player is wrong about (the same meanings for all), rather than find, for each player, a type of meaning which that player is wrong about.

8 Conclusions

People can be wrong about their own conventions; in particular, people can be wrong about the meanings of their own messages. The examples are simple, explicit, new in kind and based on an independently plausible meta-semantic story. They're artificial, in order to make the structure clear, but not fussy and not exotic. If you want to observe mistakes about conventions or meanings first-hand, then grab some friends, feed them a little misleading information, and let them play the games in the paper.

Imagine a large community of signalers who are radically wrong about the meanings of their messages. Send a field linguist among them to discover the semantics of their language using standard techniques like the Elicitation Method. The linguist will be radically misled.

Suppose Method X is taken to be a reliable test for Disease D. It's then discovered that Method X is unreliable when the subject has Condition C. Condition C is a common-or-garden condition, not involving genetic quirks or strange diets or radiation exposure. How should we react? I say: we shouldn't endorse Method X as confidently as before in cases where Condition C doesn't obtain; we should re-examine the reliability of Method X even when Condition C doesn't obtain. As with Method X and Disease D, so too with the Elicitation Method and meanings.

[23] Lewis suggests when we should interpret them as indicatives and when as imperatives. See pp. 143–7. See also [Millikan, 1995].

I don't say that we should give up the Elicitation Method; nor that the key idea behind the examples is responsible for mistakes about meanings in natural languages. I do say that if speakers can be wrong about meanings for such straightforward reasons, then we should reconsider the reliability of speakers' judgments about meanings more generally.

References

[Altshuler et al., 2019] Altshuler, D., Parsons, T., and Schwarzschild, R. (2019). *A Course in Semantics*. The MIT Press, Cambridge, MA.

[Bochnak and Matthewson, 2015] Bochnak, M. and Matthewson, L., editors (2015). *Methodologies in Semantic Fieldwork*. Oxford University Press.

[Burge, 1979] Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy*, 4(1):73–122.

[Gilbert, 1981] Gilbert, M. (1981). Game Theory And Convention. *Synthese*, 46(1):41–93.

[Gilbert, 1983] Gilbert, M. (1983). Agreements, Conventions, and Language. *Synthese*, 54(3):375–407.

[Gilbert, 1990a] Gilbert, M. (1990a). Rationality, Coordination, and Convention. *Synthese*, 84(1):1–21.

[Gilbert, 1990b] Gilbert, M. (1990b). Walking Together: A Paradigmatic Social Phenomenon. *Midwest Studies in Philosophy*, 15(1):1–14.

[Gilbert, 2013] Gilbert, M. (2013). *Social Convention Revisited*. Oxford University Press.

[Horwich, 1998] Horwich, P. (1998). *Meaning*. Oxford University Press.

[Horwich, 2005] Horwich, P. (2005). *Reflections on Meaning*. Oxford University Press, Clarendon Press ;.

[Hume, 2000] Hume, D. (2000). *A Treatise of Human Nature*. Oxford Philosophical Texts. Oxford University Press, Oxford, New York.

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

[Lewis, 1975] Lewis, D. (1975). Languages and Language. In Gunderson, K., editor, *Minnesota Studies in the Philosophy of Science*, pages 3–35. University

of Minnesota Press.

[Matthewson, 2004] Matthewson, L. (2004). On the Methodology of Semantic Fieldwork. *International Journal of American Linguistics*, 70(4):369–415.

[Millikan, 1995] Millikan, R. G. (1995). Pushmi-Pullyu Representations. *Philosophical Perspectives*, 9:185–200.

[Pinker, 1995] Pinker, S. (1995). *The Language Instinct*. Harper Perennial.

[Putnam, 1975] Putnam, H. (1975). The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science*, 7:131–193.

[Schiffer, 1972] Schiffer, S. R. (1972). *Meaning*. Oxford, Clarendon Press.

[Schiffer, 1987] Schiffer, S. R. (1987). *Remnants of Meaning*. MIT Press.

[Skyrms, 2010] Skyrms, B. (2010). *Signals: Evolution, Learning, and Information*. OUP Oxford.

[Stalnaker, 1984] Stalnaker, R. (1984). *Inquiry*. Cambridge University Press.

[Stalnaker, 1998] Stalnaker, R. (1998). Belief Revision in Games: Forward and Backward Induction. *Mathematical Social Sciences*, 36(1):31–56.

[Vanderschraaf, 1998] Vanderschraaf, P. (1998). Knowledge, Equilibrium and Convention. *Erkenntnis*, 49(3):337–369.

[Winter, 2016] Winter, Y. (2016). *Elements of Formal Semantics: An Introduction to the Mathematical Theory of Meaning in Natural Language*. Edinburgh University Press, Edinburgh.