Goal-driven Answers in the Cards Dialogue Corpus

Christopher Potts
Stanford University

Abstract

The resolvedness conditions for questions are not semantically invariant, but rather heavily influenced by the goals and intentions of the discourse participants. One of the chief innovations of recent question- and preference-driven models of pragmatics is to characterize how these high-level contextual features influence language production and comprehension. The goal of this paper is to develop and motivate general techniques for quantitatively and qualitatively assessing such models. The basis for my exploration is the Cards corpus of task-oriented dialogues, a highly structured resource that allows us not only to rigorously interpret the discourse participants' utterances but also to track their evolving goals and subgoals. The paper's central experiments provide evidence that the semantic variability of answers to *Where are you?* is governed by the particular demands of the goal that the question engages.

Proceedings of the 30th West Coast Conference on Formal Linguistics, ed. Nathan Arnett and Ryan Bennett, 1-20. Somerville, MA: Cascadilla Proceedings Project.

Cascadilla Proceedings Project
Completed July 10, 2012

Goal-driven Answers in the Cards Dialogue Corpus

Christopher Potts
Stanford University

1. Resolvedness and the task

The starting point for this paper is Ginzburg's (1995a) observation that the resolvedness conditions for questions are not semantically invariant, but rather "fixed in a particular context to a level identified by the goal" of the conversation (p. 466; see also Clark 1979; Clark & Schober 1992; Perrault & Allen 1980; Groenendijk & Stokhof 1984:II; Gibbs & Bryant 2007). Such goal-orientation is especially clear for questions like *Where are you?*, in which the appropriate level of granularity and sense of location are so highly variable; each of *my office*, *California*, *the U.S.*, *Linguistics*, and *still working on the introduction* is resolving in some contexts but non-resolving (even downright unhelpful) in others.

One of the chief innovations of models like Ginzburg's is to bring these goals (issues, plans, decision problems) into the pragmatic model, so that they can play a direct role in resolving underspecification, setting bounds on vagueness, and guiding inferences about the discourse participants' intentions. This is a leading idea of all the recent task-driven and question-driven models of contextual dynamics (Allen, 1991; Hobbs et al., 1993; Roberts, 1996, 2004; Ginzburg, 1995b, 1996; Groenendijk, 1999; Beaver, 2002; Büring, 2003; Stone et al., 2007; Beaver & Clark, 2008; Groenendijk & Roelofsen, 2009) as well as their game-theoretic and decision-theoretic counterparts (Clark, 1996; Merin, 1997, 1999; Parikh, 2000, 2001; van Rooy, 2003; Benz, 2005; Franke, 2009; Jäger, To appear; Frank & Goodman, 2012).

The goal of this paper is to develop and motivate general techniques for identifying and characterizing this kind of task dependence in real task-oriented dialogue, where the phenomena can be both quantitatively and qualitatively assessed. The basis for my exploration is the Cards corpus, a large collection of dialogues derived from a two-person collaborative game. The most noteworthy feature of this corpus for present purposes is that its transcripts record not just the dialogue exchanged by the players, but also all their actions in the game world, with high enough fidelity that we can faithfully replay the games in their entirety. Of course, we don't have access to the players' beliefs and intentions, but we do have enough linguistic and behavioral cues to make confident inferences about them.

My focus is on the corpus's 800+ answers to the (implicit or explicit) question *Where are you?* Even in the highly constrained Cards world, these answers come at numerous levels of granularity, from *middle of the board* to *2 spaces to the left of the gap underneath the entrance to the middle room*. Intuitively, the preferred granularity is governed by the players' subgoal at that point in the game. For example, where they want to divide up the board for general exploration, very general answers are preferred. When they need to meet or find specific things, only highly specific answers are resolving.

After describing the Cards corpus in some detail (sec. 2), I look more closely at the nature of these locative answers (sec. 3), and then begin the process of setting up an experiment to test the hypothesis of goal-orientation. This involves some semantic annotation (sec. 4) and, more interestingly, a technique for interpreting these annotations in the Cards world itself (sec. 5). With these preliminaries in place, it is straightforward to define a precise notion of granularity and show that it correlates with the players' subgoals (sec. 6). The paper closes with a more precise theoretical explanation for this correlation (sec. 7) and some suggestions for additional experiments along these lines that could be conducted fairly easily

* Thanks to Pranav Anand, Mike Frank, Noah Goodman, Jesse Harris, Dan Lassiter, Sven Lauer, Adam Vogel, and audiences at WCCFL 30, NYU, and the University of Rochester for valuable comments, suggestions, and discussion. Thanks also to Alex Djalali, Sven Lauer, Karl Schultz, and the whole SUBTLE team for help with corpus collection. This research was supported in part by ONR grant No. N00014-10-1-0109 and ARO grant No. W911NF-07-1-0216.

Cascadilla Proceedings Project
Completed July 10, 2012

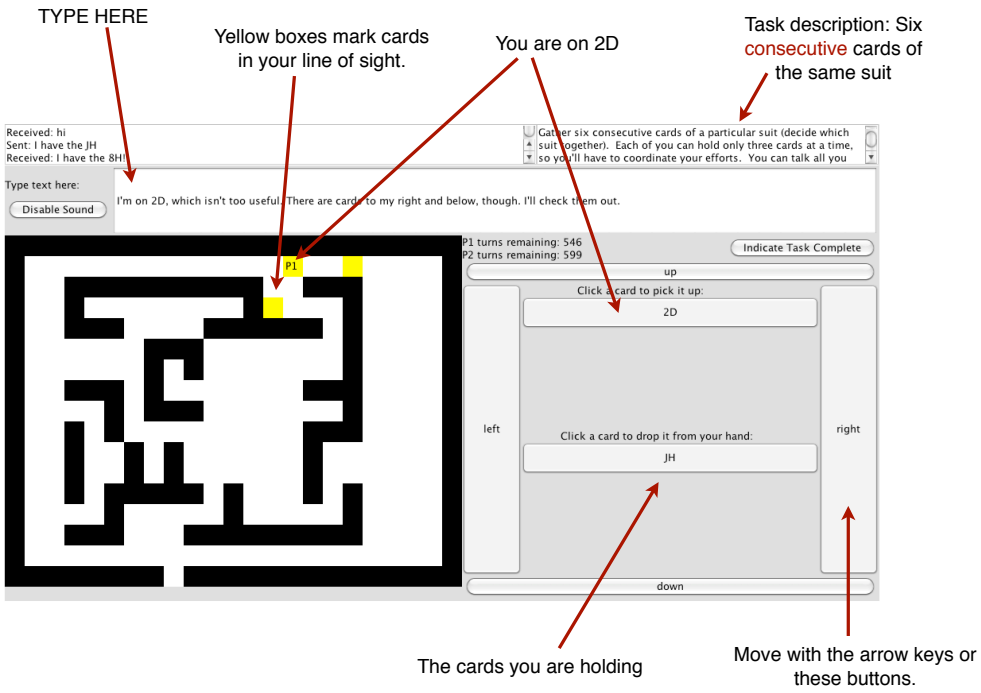


Figure 1: An annotated version of the Cards gameboard.

with the Cards corpus (sec. 8).

The corpus is available at `http://cardscorpus.christopherpotts.net/`. The distribution includes the transcripts and starter code for working with them in Python and R, and the site provides a search and visualization function. I have also posted the annotations and R functions for extracting and visualizing the denotations used in this paper, and for running the experiments of sec. 6.

2. The Cards corpus

The Cards corpus was designed to engage models in which language production and comprehension are driven by the discourse participants' goals (issues, plans, questions, decision problems). The corpus consists of 1,266 transcripts from a Web-based, two-person collaborative search game. The game-world is a maze-like environment in which a deck of 52 playing cards has been randomly distributed. When players sign on, they are shown the annotated gameboard in fig. 1 to help orient them, and then they are presented with the following task description, which also remains visible in the upper-right corner of the screen throughout play:

(1) Gather six consecutive cards of a particular suit (decide which suit together), or determine that this is impossible. Each of you can hold only three cards at a time, so you'll have to coordinate your efforts. You can talk all you want, but you can make only a limited number of moves.

This task is intentionally underspecified. The players are thus forced to negotiate a specific goal and then achieve it together. In terms of question-driven models (Ginzburg, 1996; Roberts, 1996; Büring, 2003), the task establishes a hierarchy of questions: 'What is going on?' (as the players get used to the game world and its somewhat unfamiliar objects), then 'Which suit should we pursue?', then 'Which sequence?', and finally 'Where is card X' for relevant values of X, with relevance defined by how the more general questions were resolved. Because of the highly structured nature of the game and its domain of objects, these questions and their relationships can all be modeled precisely.

Experienced players have already resolved 'What is going on?', so they tend to skip directly to 'Which suit?', which they handle by first exploring a little bit and then reporting on their findings. The exchange in (2) is a typical opening for experienced players. (Between utterances, the players explore

the environment and manipulate cards; this dialogue spans a total of 80 moves. Player 1's orthographic ambiguity on line 5 is amusing.)

(2) P2: hi
P1: hello
P1: what suit we going with?
P2: What's around you?
P1: Ks and Qs
P2: of what?
P1: king of spades and queen of spades
P2: ok let's go with those then
P2: I'll start looking for high spades
P1: ok..

In order to stimulate conversation and genuine collaboration, we impose a number of constraints on the players' actions, and we partly or totally conceal important information about the game world:

(3) a. Each player can see her own location (a yellow box containing "P1" or "P2", depending on the order in which the players signed on), but she cannot see the location of the other player. This pushes the players to describe their locations for purposes of finding each other, locating cards, and formulating search strategies.

b. The cards are visible to a player only when they are in her "line of sight", which is typically three squares. This encourages the players to adopt flexible plans and revise them based on what they find during exploration.

c. Each player can hold at most three cards. This means that no single player can complete the task alone. Combined with the above constraints, it stimulates conversation about the locations of cards and the players' current holdings.

d. Each player has a limited number of moves, where a move is any action in the game except chatting. This provides an incentive to formulate strategic plans as opposed to exhaustively exploring the space. It also means that the players are free to discuss their plans and formulate new ones, which they do especially when they start to run low on moves.

When the players are finished, they click TASK_COMPLETE and the transcript is written to a file. Each transcript records not only the chat history, but also the initial state of the environment and all the players' actions (with timing information) throughout the game, which permits us to replay the games with perfect fidelity. In all, the corpus contains 45,805 utterances (mean length: 6 words), totaling 282,065 words, with a vocabulary size around 4,700. Most actions are not utterances, though: there are 371,811 movements, 19,157 card pick-ups, and 12,325 card drops. The median game length is 373 actions, though this is extremely variable (standard deviation: 215 actions). The mean game length in minutes is 8:30 (standard deviation: 5:22).

The transcripts are in CSV format. Tab. 1(a) is an example of the high-level environmental information included in the files, and tab. 1(b) is a snippet of gameplay. By stepping through the transcripts line-by-line, one can establish what the initial context is like and then update it using the content from each line/event. Thus, each transcript gives rise to a sequence of ⟨context, event⟩ pairs. Computationally, we truly have language in context, which opens up new opportunities for studying not only how the context shapes the players' language, but also how their language shapes the context.

The corpus was collected in a somewhat naturalistic fashion; it took us a long time and a number of iterations to find settings for the game that encouraged meaningful communication. As a result, there are points of variation that users of the corpus should be aware of.

Figs. 2(a)–2(d) depicts the four gameboards that appear in the corpus. The gray squares are walls and the black squares are invisible walls, which appear to players only when within their line of sight. Our most heavily used configuration is the one in fig. 2(a), which is complex enough that it is worth the players' time to formulate a search plan but not so complex that they become frustrated. In contrast, the large size of tab. 2(d) and the unpredictability of figs. 2(b)–2(c) tend to stump novice players. Fig. 2(b)

Agent	Time	Action type	Contents
Server	0	COLLECTION_SITE	Amazon Mechanical Turk
Server	0	TASK_COMPLETED	2010-06-17 10:10:53 EDT
Server	0	PLAYER_1	A00048
Server	0	PLAYER_2	A00069
Server	2	P1_MAX_LINEOFSIGHT	3
Server	2	P2_MAX_LINEOFSIGHT	3
Server	2	P1_MAX_CARDS	3
Server	2	P2_MAX_CARDS	3
Server	2	P1_MAX_TURNS	200
Server	2	P2_MAX_TURNS	200
Server	2	GOAL_DESCRIPTION	Gather six consecutive cards …
Server	2	CREATE_ENVIRONMENT	[ASCII representation]
Player 1	2092	PLAYER_INITIAL_LOCATION	16,15
Player 2	2732	PLAYER_INITIAL_LOCATION	9,10

(a) Environment metadata in the corpus format.

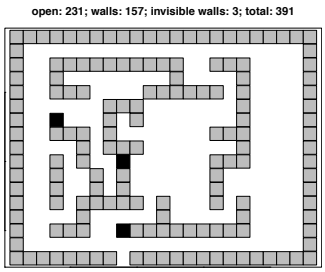
Agent	Time	Action type	Contents
Player 1	494994	CHAT_MESSAGE_PREFIX	ok
Player 2	495957	CHAT_MESSAGE_PREFIX	i found the 2c. where are you at
Player 1	499620	PLAYER_MOVE	5,14
Player 1	500708	PLAYER_MOVE	5,13
Player 1	503163	PLAYER_MOVE	5,12
Player 1	523339	CHAT_MESSAGE_PREFIX	i got 2c so i drop the 10c?
Player 2	533955	CHAT_MESSAGE_PREFIX	yes
Player 1	539971	PLAYER_DROP_CARD	5,12:10C
Player 1	542119	PLAYER_PICKUP_CARD	5,12:2C
Player 1	552576	CHAT_MESSAGE_PREFIX	got it now what

(b) A snippet of gameplay in the corpus format.

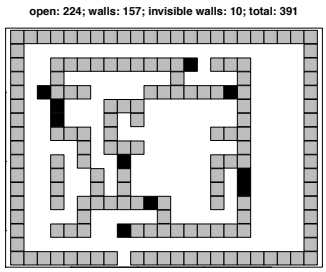
Table 1: The CSV format of the corpus release.

is particularly devious and thus particularly exasperating for players, because it is potentially infeasible. The room on the left of the middle is (invisibly) walled off. Crucial cards might be trapped in this zone with both players outside of it, or else one player might be trapped in this zone with too few useful cards available. This is why we have the parenthetical "determine that this is impossible" in the task description (1). However, as suggested by the small number of transcripts from this variant, we ultimately decided that this was not the most effective way to stimulate rich interactions.

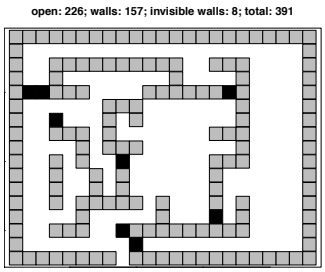
Tightly controlling the maximum number of (non-utterance) moves each player could make turned out to be a more reliable method for eliciting interesting dialogues and strategies. Fig. 2(e) summarizes the variation here. The first four rows involve symmetric conditions: each player has the same number of turns. The final three rows are radically asymmetric: one player is barely able to move at all, while the other has a high number of moves. In these asymmetric conditions, the player with relatively few moves has an unlimited line of sight: she can see where there are cards, but not their identities. The more mobile player has the usual three-squares line of sight. In these games, the less mobile player typically remains stationary and guides the other player to card-rich areas of the board. I like to think of this as a kind of advisor–advisee situation. The less mobile advisor has a broad but incomplete view of the landscape. The more energetic but less knowledgeable advisee does the hard work of exploration and



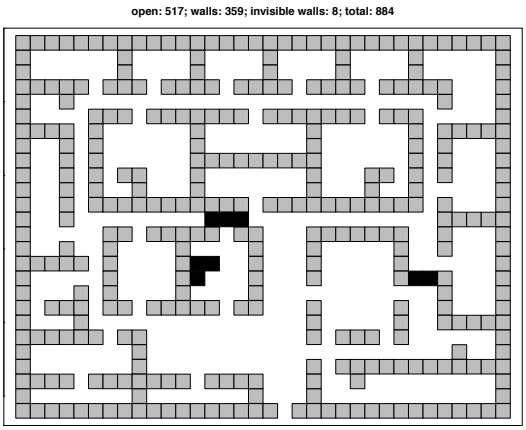
(a) 977 transcripts.



(b) 52 transcripts.



(c) 144 transcripts.



(d) 93 transcripts.

Type	Player 1	Player 2	Transcripts
Symmetric	100	100	267
	140	140	4
	500	500	137
	600	600	393
Asymmetric	40	700	76
	40	500	50
	40	350	339

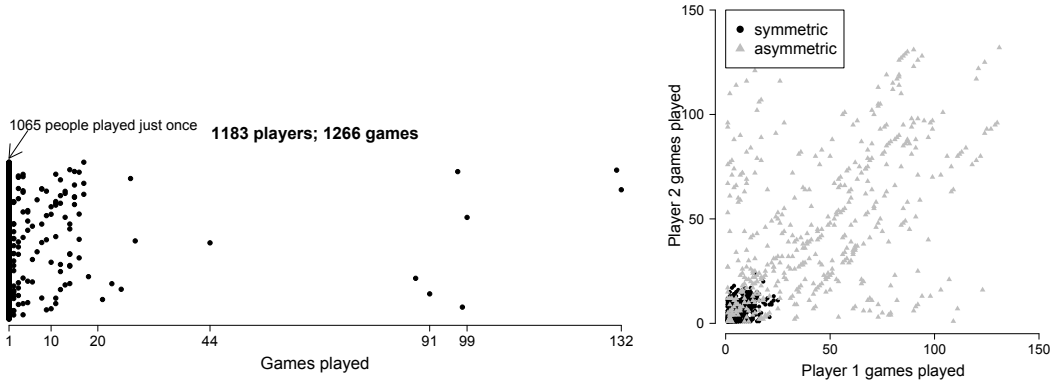
(e) The distribution of maximum allowed moves for each player.

Figure 2: Noteworthy points of variation in the corpus: map types and maximum non-utterance moves. The map titles provide counts of the different square types. Walls visible only within the line of sight are in black.

brings his findings back to the advisor. In my view, the best transcripts are the asymmetric ones in which the mobile Player 2 has 350 moves and the symmetric ones where the move counts are relatively low (100 or 140), so that the players feel a real pressure to be strategic.

Most of the players were recruited via Amazon's Mechanical Turk. This made it relatively easy to recruit large numbers of people to play the game with each other in real time (release 2 involves games from 1,183 different players). It also allowed us to include (anonymized versions of) the Amazon Worker Ids in the transcripts (tab. 1(a), lines 3-4), which means that the corpus supports work on how novice strategies differ from expert ones, how experts accommodate novices, and so forth (Djalali et al., 2011a).

Another advantage of crowdsourcing was that the players typically did not know each other outside of playing the game. Chat is notorious for being hard for outsiders to understand, in large part because it is often conducted between close acquaintances who can rely on very rich common ground and thus produce short, elliptical messages. Our players generally met through the game, so they had only the game context to fall back on, which makes the dialogues more amenable to annotation and modeling. This does not rule out the possibility of studying the effects of rich common ground, though. Some players played so often that they developed common ground of their own, and this had a measurable impact on their play (Djalali et al., 2011a). Fig. 3 provides information on how often people played and what the pairings were typically like. The important thing about fig. 3(b) is that, for both symmetric and asymmetric conditions, we have a wide variety of different combinations of novices and experts. There



(a) Number of games played by each player. The data points have been jittered randomly along the y-axis to make their clustering evident.

(b) Pairings.

Figure 3: Information about the players.

is less variation in the symmetric condition due to the contingencies of data collection, but even there we have many different pairings of novice and expert players. (The game is not hard;[1] after a few rounds, most players have developed styles and general strategies.)

As I mentioned above, the Cards corpus and associated computational tools can be freely downloaded from `http://cardscorpus.christopherpotts.net/`. Its chief selling point, to my mind, is that it is large enough to support quantitative work and yet structured enough to permit researchers to isolate very specific phenomena and make confident inferences about the participants' intentions. While there are a number of other excellent task-oriented corpora available (Thompson et al., 1993; Allen et al., 1996; Stoia et al., 2008; Blaylock & Allen, 2005), as far as I know, only Cards has this particular mix of properties.

3. *Where are you?* and its responses

Djalali et al. (2011b) use version 1 of the Cards corpus to explore the ways in which the players' evolving approach to the central objective ('six consecutive cards of the same suit') affects their use and interpretation of underspecified referential expressions and quantifiers. Djalali et al. show that the task itself gives rise to a notion of relevance that is a powerful predictor of both where speakers feel they can get away with using underspecified expressions and how those expressions will be interpreted.

I take a similar approach in this paper, but my focus is not on the central objective of the game, but rather on a class of subgoals that it gives rise to, namely, those involving the locations of players and cards. These subgoals are well-suited to investigation for a mix of practical and theoretical reasons. On the practical side, the gameboard is a matrix of discrete squares, so the space of possible locations is precisely defined and tractably small (between 224 and 517 $\langle i, j \rangle$ coordinates; see fig. 2). In addition, all this locational information is encoded in the transcripts — the random initial positions of the cards and players are included directly, and from there one just follows the players as they travel and manipulate the cards. On the theoretical side, locational information is observed by only one player (see (3)), which means that it becomes mutually and publicly known only via dialogue. It is also absolutely crucial information for effectively searching the space and accumulating a winning hand, so it is always relevant.

My primary data are the answers to implicit and explicit questions about the players' locations. In the corpus as a whole, I have identified 690 explicit instances of one player asking another *Where are*

[1] Well, it isn't hard, but it should be noted that only 35% of the games are actually winning games, even when we remove the 'infeasible' condition. A high percentage of players seem to *believe* they succeeded, when in fact they got confused about their holdings, the success conditions, or both.

you?, sometimes with modifiers like *exactly* and sometimes with informal spellings, mis-spellings, and so forth. Alongside these are 849 answers. (The number of answers is larger than the number of questions because some answers are to implicit questions and some questions receive multiple responses.)

As I said, the game world is a grid. One can therefore always give a very precise location, by counting the number of grid squares from a precise reference point. And, indeed, some players do, on some occasions, identify their locations this way. However, such answers are rare and represent just one of a wide variety of general answer types. Here is a random sample organized by length of utterance, which positively correlates with specificity:

(4)
a. bottom
b. bottom left
c. left middle
d. I'm dead center
e. in the middleish
f. i'm inside the dead-end hallway
g. in the small box at the left center
h. I am in the middle toward the bottom.
i. far right, 7 blocks up from the bottom
j. I am in the middle just under the C room
k. i'm in the narrow room in the upper left
l. im inside the sideways C at the top left
m. The bottom left corner above the second line
n. I am two squares away from the upper left corner
o. I am just to the left of the C room in the middle.
p. I am in the long rectangle towards the bottom center
q. I am 2 spaces off the top 3 from left wall in the center
r. bottom right corner inside the box just below the single black square
s. i am at the very bottom right in front of the long skinny corridor across the bottom

Most of the answers at the very top of (4) denote large subsections of the gameboard: *middle*, *bottom left*. They also tend to be vague by design (*middleish*) or because of ambiguity of reference (*small box*, *C room*). In contrast, the answers at the bottom of the list describe very specific regions of the gameboard. They are still vague in their own ways, but the indeterminacy is comparatively low.

The pragmatic theories I mentioned in the introduction lead us to expect that this variation will be tightly controlled by the high-level goals of the discourse participants. And, indeed, when one reads through the transcripts looking at these answers in context, it seems clear that the level of specificity is task dependent. For example, when the players are establishing a basic search strategy, their answers tend to pick out large regions of the board. The exchange in (5) illustrates. Player 1 poses the initial *where are you*, both players reply by naming quadrants of the board, and then Player 2 uses this as the basis for proposing an equitable search strategy.

(5)
P1: Hey
P2: hi
P1: Where are you?
P2: i am upper left
P1: Okay good I'm upper right
P2: split down the middle top to bottom?
P1: sounds good

The above is from the very start of a transcript. Search tasks also arise later on, when the players need specific cards that they have yet to locate:

(6) P1: so we need 9
P2: I guess so
P1: where are you now?
P2: Im near bottom right corner now. You?
P1: in the middle room
P1: nothing up the right wall, already tried there

In contrast, when the players need to find each other to exchange a card, they are more specific:

(7) P2: ok i am going to drop a card for you
P1: ok, where at?
P2: i am at the very bottom right in front of the long skinny corridor across the bottom
P2: right at the opening on the left side
P1: ok

This level of specificity is also typical of situations in which one player has located a needed card but is already holding three cards and thus needs the other player to come pick it up:

(8) P2: I found 6S
P2: but I can't pick it up
P1: Where are you?
P2: near the top
P2: where the one opening is to go down

Though there is a grid, the player seem to find it clumsy, unreliable (there are no labels on the axes), or unnecessary (perhaps due to the line-of-sight horizon). As a result, in situations in which a precise grid square is at issue, they generally use descriptions like those seen in (7) and (8). Since these are often vague, it is common for the players to engage in what Clark & Wilkes-Gibbs (1986) call *collaborative reference*, in which the two players work together to ensure coordination on the relevant point (see also DeVault & Stone 2007). In (9), for example, Player 1 seems to find Player 2's initial answer insufficient and so asks for clarification.

(9) P1: where are you
P2: middle
P1: I found 4c
P2: 4 from the u
P1: where in the middle exactly
P2: to the right of the u
P1: the u under the c

Similarly, in (10), the initial *Where are you?* is first met with another question, which serves to make sure that the two players have the same referent in mind for the expression that Player 2 want to use to anchor his locative description.

(10) P1: Where are you?
P2: you see the c in the middle
P1: yeah
P2: I;m to the left of it

Example (11) is striking because of how explicit the players are about the goals and because they are admirably strategic about finding the least-cost way to achieve them.

(11) P2: where are you so I know where to drop cards off?
P1: i am just on the outside of the center box on the back on the c looking box on the left side of center
P1: know where i am talking about?
P1: i can always move
P2: well go to the top left corner of the outside then and wait I guess?

The above observations suggest the following hypotheses:

(12) **Specificity hypotheses (informal versions)**:

a. When the players need to meet up or direct each other to specific cards, their answers will tend to be more specific.

b. When the players are developing a general search strategy, their answers will tend to be less specific.

Hypothesis (12a) is immediately intuitive. If I want you to move to a specific square, a general answer like *right side* has a very low probability of success. Hypothesis (12b) is slightly less obvious, since a specific location like *very bottom, two squares from the right wall* would suggest the same search strategy as *bottom right*. However, that level of specificity is not needed, so we can expect speakers to avoid the unnecessary effort. In addition, such specificity might suggest confusion about the task itself, which provides further reason to avoid it. (I discuss this further in sec. 7.)

The remainder of this paper is devoted to making these hypotheses more precise and testing them. My general strategy is to develop a data-driven notion of specificity for locative expressions and then assess how well it correlates with the players' immediate subgoal. The semantic annotations I describe next are the foundation for this work.

4. Annotations

To begin refining the above hypotheses, I annotated all of the locative question–answer pairs in the subset of the data involving the map in fig. 2(a). It would be straightforward to extend the annotations to transcripts based on the other maps, but merging the results for experiments would be challenging, since the walls, landmarks, and boundaries would no longer be constant across the data set.

4.1. Questions

The question annotations bracket the question, provide a simple semantic parse (`sem=`), and, most importantly, indicate which goal the question engages (`engagesGoal=`). Some typical examples are given in tab. 2(a). The questions are mostly of the form *where are you?*, including some mis-spellings, some involving adverbs like *exactly*, restrictions like *where in the middle are you?*, and a handful of polar interrogatives that I felt compelled to include in order to capture the full interaction. In all, I annotated 535 overt questions and identified an additional 64 implicit questions as part of annotating locative answers (see the next subsection).

The most important variable in the question annotations is `engagesGoal=`, which specifies exactly one of four goal types:

(13) a. `search`: the players are exploring the space looking for cards whose precise locations they do not know.

b. `card`: one player knows the location of a specific card and is trying to convey that location to another player.

c. `meet`: one player is trying to get the other player to occupy the same space as she does.

d. `?`: where I couldn't determine what was going on.

Tab. 3 provides summary information about these goal annotations. I've given the counts relative to the symmetric–asymmetric distinction (fig. 2(e)), which reveals that asymmetric games are notably more likely to generate `meet` tasks in virtue of the fact that, as I mentioned in sec. 2, the expert strategy in such games is for Player 1 to retreat to a specific location and then instruct the more mobile Player 2 on finding cards and bringing them to Player 1.

Annotating for the goal was easier than I expected it to be. For search conditions, the question is often followed quickly by a purpose clause like *so we don't overlap*, an overt suggestion like *Let's split the maze*, or another sort of explicit suggestion, as seen in (5). Similarly, in `card` and `meet` conditions, the discourse is typically centered around a specific card, with the players' goals and plans spelled out explicitly in the dialogue. Moreover, given the focussed nature of the games, there are not a whole lot

```
Player 2,23804,CHAT_MESSAGE_PREFIX,[where are you?]_{sem=where(you);
engagesGoal=search}

Player 2,20236,CHAT_MESSAGE_PREFIX,hello, [where are you
located?]_{sem=where(you); engagesGoal=search}

Player 2,13931,CHAT_MESSAGE_PREFIX,[where are you
P?]_{sem=where(you); engagesGoal=meet}

Player 1,204774,CHAT_MESSAGE_PREFIX,[where are you
exactly]_{sem=exactly(where(you)); engagesGoal=card}
```

(a) Questions.

```
Player 2,31931,CHAT_MESSAGE_PREFIX,I am [on the left side in the
middle]_{sem=located(Player 2; @left middle); answers=23804}

Player 1,69058,CHAT_MESSAGE_PREFIX,hi....i am [in the lower
right corner of the center of the board]_{sem=located(Player 1;
@middle<bottom right>)); answers=20236}

Player 1,22344,CHAT_MESSAGE_PREFIX, i am [on the left bottom
corner]_{sem=located(Player 1; @bottom left corner); answers=13931}

i'm [at the very very top right]_{sem=located(Player 1; @precise top
right); answers=imp(where(you)); engagesGoal=card}
```

(b) Answers.

Table 2: Annotations.

of other reasons for players to mention locations, which is why there are relatively few `?` tokens. (For the most part, I suspected that `?` situations were `search` ones but couldn't tell because the surrounding dialogue was sparse.)

4.2. Answers

The annotations for answers are more involved than those for questions. Some examples are given in tab. 2(b). The locative part of the utterance is in square brackets. The value of `sem=` is the object of predication (either Player 1 or Player 2) and then, after the `@` sign, a semantic representation, which is a simplified version of the bracketed raw text. (Where I could not figure out how to simplify it, only an `@` appears and the example was left out of the experiments below.)

In the semantics, the predicates are separated by semicolons. I interpret these conjunctively in this paper, but the order of the predicates would support treating *precise* and *approx*, which pool a wide array of adverbials, as modifying only the predicates to their right. The one piece of hierarchical semantic structure I do pay attention to involves the `middle<bottom right>` notation of the second example in tab. 2(b). Here, the initial predicate provides an explicit domain in which to interpret the predicates in angled brackets. In the text, these usually correspond to partitive phrases like *middle of the board*.

The value of `answers=` is the timing value for an annotated *where are you?* question earlier in the discourse. In such cases, the utterance inherits the `engagesGoal=` value from the question. Where there is no such question, `answers=` has `imp(where(you))` as its value and an `engagesGoal=` value is provided directly on the answer, as in the final example in tab. 2(b).

In adding the semantic representations, I worked hard to keep the overall vocabulary small. The steps I took included (i) collapsing together variant ways of referring to certain regions like *middle* and *center*, *reverse C* and *backwards C*; (ii) pooling modifiers like *very*, *exactly* and *all the way* into a single

	Symmetric	Asymmetric	Total
`?`	3	6	9
`card`	35	2	37
`meet`	22	218	240
`search`	183	130	313
Total	243	356	599

Table 3: The distribution of `engagesGoal` annotations.

predicate *precise* (and similarly for hedges and approximators, which become *approx*); and (iii) ignoring a lot of filler words and syntactic variation. It would be preferable to make more direct use of the raw text, but the present perspective yields a clearer picture of the players' referential strategies. (For richer logical and linguistic models that might be adaptable to this domain, see Kress-Gazit & Pappas 2010; Finucane et al. 2010; Tellex & Roy 2009; Tellex 2010; Golland et al. 2010.)

5. Semantics for locative expressions

In order to address the specificity hypotheses (12), we have to interpret players' answers in terms of the Cards world. The annotations described above provide shallow semantic representations, so we can use those as the basis for interpretation. To keep things simple, I additionally strip off prepositional modifiers like *above* in *above the C room*. The resulting complete lexicon is given in tab. 4 along with the token counts for each item. SQUARE pools all the mentions of specific squares by their coordinates or by counting from a landmark, and BOARD is used exclusively as a domain restriction.

Word	Count	Word	Count	Word	Count	Word	Count
BOARD	547	corner	91	hall	31	U_room	2
right	227	approx	77	room	18	T_room	2
middle	195	SQUARE	71	sideways_C	11	deadend	2
top	183	precise	68	loop	7	wall	1
left	178	entrance	59	reverse_C	3	sideways_F	1
bottom	169	C_room	35				

Table 4: Semantic lexicon with token counts.

The *well-formed formulae* are conjunctions of lexical items with a domain specification:

Definition 1 (Well-formed formulae)**.** The set of well-formed formulae is the set of all expressions $\delta(\varphi_1 \wedge \cdots \wedge \varphi_k)$ where δ and all the φ_x are items from the lexicon.

The domain is usually the whole board, in which case it is given as BOARD. However, as I mentioned briefly when describing the annotations, there are sometimes non-trivial domain specifications: 47 in which the domain is *middle*, 3 in which it is *C_room*, and 1 in which it is *hall*. The corresponding utterances are things like *top right of the middle* or *on the bottom in the inner part of the board*. Semantically, the domain specification just serves to distinguish, for example, *top of the middle* from *top* (interpreted as *top of the board*).

The full domain of interpretation for these locative expressions is the gameboard, which I treat as a matrix with two kinds of cells: walls and open areas.

Definition 2 (Gameboard)**.** A Cards gameboard is a matrix $G_{m\times n}$ where m is the number of rows and n is the number of columns. A subset of the cells in $G_{m\times n}$ are walls, which neither players nor cards can occupy.

We can now begin building semantic interpretations. It is tempting at first to rely solely on intuitions. The lexicon is fairly small, so it seems at first that we could go through it and specify, for each item,

which subset of the squares it denoted. The word *top* would refer to the top half, *right* to the right half, *top right* to their intersection, and so forth. However, if ones tries this, one is immediately hampered by the vagueness of these phrases. It is hard or impossible to determine where the top ends and the middle begins, what it means to be near the C room, how *very bottom* differs from *bottom*, and on and on — only the expressions identifying specific squares yield easy answers.

I address this by extrapolating the necessary interpretations from the corpus itself. The basic building block for this is the *phrase location matrix*, which, relative to a gameboard $G_{m\times n}$, maps each formula ψ to the count matrix C in which each cell $C[i,j]$ contains the number of times that a player predicated ψ of himself while standing on $\langle i,j\rangle$:

Definition 3 (Phrase location matrices)**.** Let $G_{m\times n}$ be a gameboard and $\delta(\varphi_1\wedge\cdots\wedge\varphi_k))$ a well-formed formula. Then $Loc_{G_{m\times n}}(\delta(\varphi_1\wedge\cdots\wedge\varphi_k))$ is the $(m\times n)$ matrix C such that, for all i,j,

$$C[i,j]=\begin{cases}\text{if } C[i,j] \text{ is a wall in } G_{m\times n}\text{: } \textit{undefined}\\ \text{if } i<0 \text{ or } i\geqslant m\text{, } \textit{undefined}\\ \text{if } j<0 \text{ or } j\geqslant n\text{, } \textit{undefined}\\ \text{otherwise: the number of times that all of the expressions } \varphi_x \text{ appeared}\\ \quad\text{together as part of an expression used by some player at } \langle i,j\rangle\\ \quad\text{with domain restriction } \delta\end{cases}$$

A partial phrase location matrix for $\textsc{board}(\textbf{top}\wedge\textbf{right})$ is given in tab. 5(a), using the gameboard in fig. 2(a). The empty spaces are undefined values (walls). The high counts are all in the top right, which is reassuring. However, because of the relatively small size of the data set, there are odd gaps, and the outliers (e.g., the 1 that is four rows from the bottom) look like they will have too much sway.

To address this, I use a simple smoothing function, which adjusts each cell in a phrase location matrix using the values of the non-wall cells above, below, to the left, and to the right of it:

Definition 4 (Smoothing)**.** Given an $(m\times n)$ phrase location matrix C, the smoothing of C, written $Smooth(C)$, is a new matrix S with the same dimensions as C such that, for all $0\leqslant i<m$ and $0\leqslant j<n$,

$$S[i,j]=\begin{cases}\text{if } C[i,j]=\textit{undefined}\text{: } \textit{undefined}\\ \text{otherwise: } \frac{2C[i,j]+\sum_{x\in neighbors}x}{|neighbors|}\end{cases}$$

where *neighbors* is the vector $\big[C[i{-}1,j],C[i{+}1,j],C[i,j{-}1],C[i,j{+}1]\big]$ with all undefined values removed and $|neighbors|$ is its number of elements.

When this smoothing algorithm is applied repeatedly, the gaps are filled in and the outliers are diminished. Tab. 5(b) shows the results of smoothing tab. 5(a) 10 successive times. The outlier four rows from the bottom basically disappeared, and the top right is denser. The meaning is still vague, but respectably so, with the counts getting steadily lower as we travel outward from the very top right.

The formulae in our corpus have very different overall frequencies. To abstract away from this point of variation, I normalize the smoothed matrices by dividing by their total size:

Definition 5 (Probability matrices)**.** Let M be an $(m\times n)$ matrix (of raw counts or smoothed values), and let T be the sum of all the values in M. Then F is the $(m\times n)$ matrix such that, for all $0\leqslant i<m$ and $0\leqslant j<n$,

$$F[i,j]=\frac{M[i,j]}{T}$$

Putting the above definitions together, we arrive at an interpretation function that is parameterized by the gameboard $G_{m\times n}$ and the number of times s that we smooth the count matrix:

Definition 6 (Interpretation)**.** Let ψ be a well-formed formula. The interpretation of ψ in gameboard $G_{m\times n}$ relative to smoothing factor s, written $[\![\psi]\!]^{G_{m\times n},s}$, is the probability matrix F such that, for all $0\leqslant i<m$ and $0\leqslant j<n$,

$$F[i,j]=\frac{S}{T}$$

where $S=Smooth_s(\cdots(Smooth_1(Loc(\psi))))$ and T is the sum of all the values in S. Setting s to 0 means no smoothing.

Because the denotations are probability distributions, we can use their information-theoretic entropy as a measure of their specificity:

Definition 7 (Entropy)**.** Let ψ be a formula and $F = [\![\psi]\!]^{G_{m\times n},s}$ be its interpretation relative to gameboard $G_{m\times n}$ and smoothing factor s. The entropy of ψ is

$$H(F) \stackrel{def}{=} -\sum_{i=0,j=0}^{m,n} (F[i,j] \cdot \log_2 F[i,j])$$

where $0 \cdot \log_2 0 = 0$ (Cover & Thomas 1991: §2).

Tab. 5(c) shows the probability distribution derived from tab. 5(b). These matrices are my proposal for probabilistic (vague) denotations for locative formulae in the Cards corpus. The numerical depiction in tab. 5(c) is not especially perspicuous. The nature of these denotations is more immediately evident in heatmap representations like those in fig. 4, in which darker colors correspond to higher probability. For all these pictures, 10 rounds of smoothing were done. The entropy values (given in the plot titles) correlate well with the intuitive notion of specificity of reference.

6 4 2 3 5 7	2.30 2.54 2.51 2.54 2.61 2.65	0.035 0.038 0.038 0.038 0.039 0.040
3 0 1 1	2.27 2.31 2.40 2.45	0.034 0.035 0.036 0.037
2 3 0 4 3	0.89 1.93 1.97 2.04 2.09	0.013 0.029 0.030 0.031 0.031
1 1 2 0 1	0.67 1.57 1.60 1.65 1.69	0.010 0.024 0.024 0.025 0.025
0 2 1 4 1	0.36 1.22 1.24 1.26 1.28	0.005 0.018 0.019 0.019 0.019
0 1 0 2 0	0.24 0.85 0.86 0.88 0.89	0.004 0.013 0.013 0.013 0.013
0 0 0 0	0.52 0.52 0.53 0.54	0.008 0.008 0.008 0.008
... 0 0 0 0 0	... 0.00 0.28 0.27 0.28 0.29	... 0.000 0.004 0.004 0.004 0.004
0 0 0 0	0.14 0.14 0.15 0.16	0.002 0.002 0.002 0.002
0 0 0 0 0 0	0.03 0.04 0.08 0.10 0.11 0.12	0.000 0.001 0.001 0.002 0.002 0.002
0 0 0 0 1 0	0.02 0.05 0.08 0.10 0.11 0.11	0.000 0.001 0.001 0.002 0.002 0.002
0 0 1 0 0	0.01 0.10 0.10 0.11 0.11	0.000 0.001 0.002 0.002 0.002
0 0 0 0 0	0.00 0.08 0.08 0.08 0.08	0.000 0.001 0.001 0.001 0.001
0 0 0 0	0.05 0.05 0.05 0.05	0.001 0.001 0.001 0.001
0 0 0 0 0 0	0.00 0.01 0.03 0.03 0.03 0.03	0.000 0.000 0.000 0.000 0.000 0.000
(a) Counts.	(b) Smoothed.	(c) Probabilities.

Table 5: The rightmost seven columns of the gameboard, with the counts for **BOARD**(**top** $\wedge$ **right**). The blank spaces correspond to walls (undefined values). The counts table shows an outliner four rows from the bottom. Smoothing greatly reduces its value compared to the 1s that are in the top right among higher values. The third matrix shows the associated probability distribution, which is the lexical denotation for this formula and also the basis for the entropy measurement.

6. Experiment

We're now in a position to refine and test the hypotheses in (12), which I repeat here:

(12) **Specificity hypotheses (informal versions; repeated)**:

a. When the players need to meet up or direct each other to specific cards, their answers will tend to be more specific.

b. When the players are developing a general search strategy, their answers will tend to be less specific.

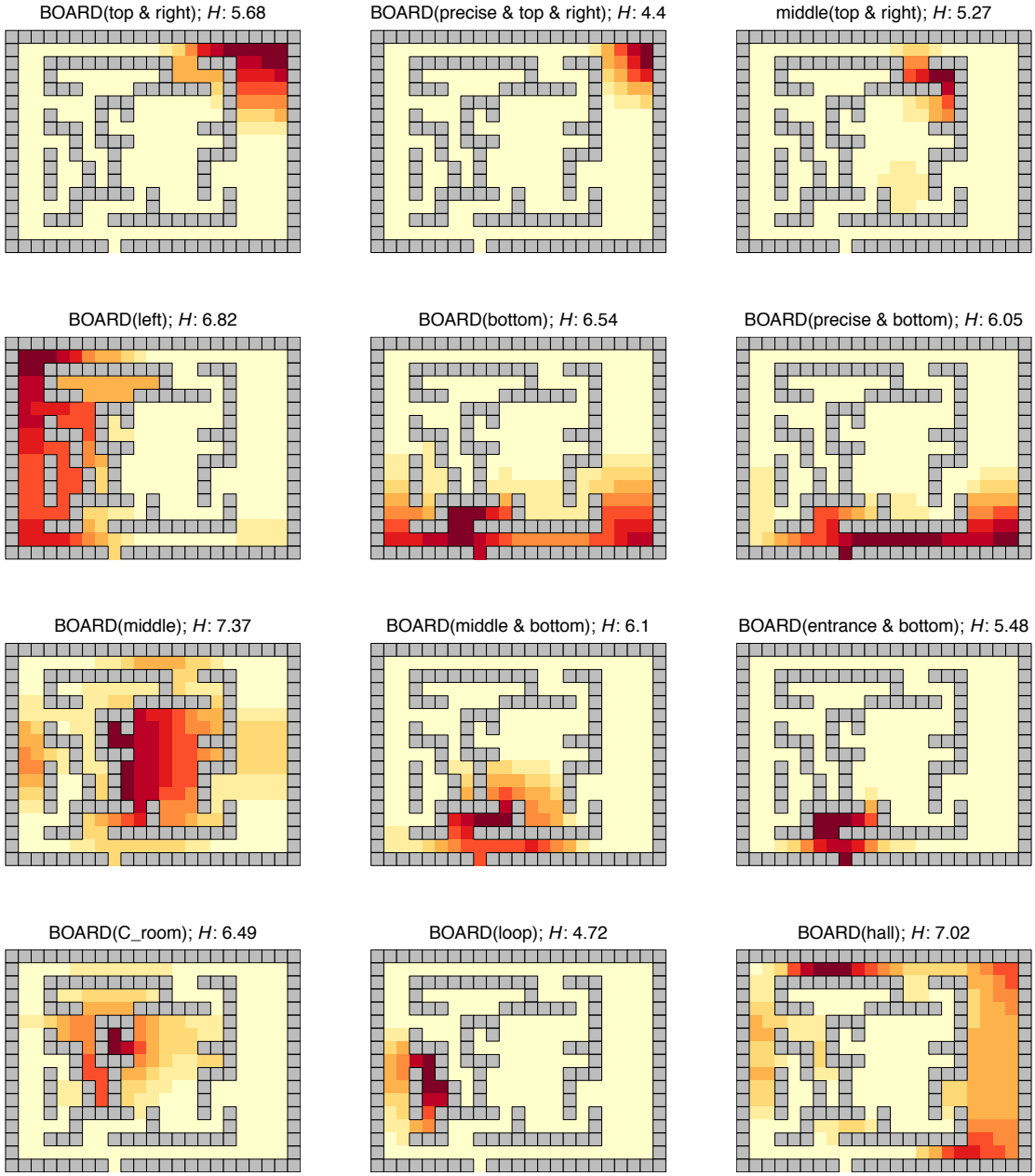
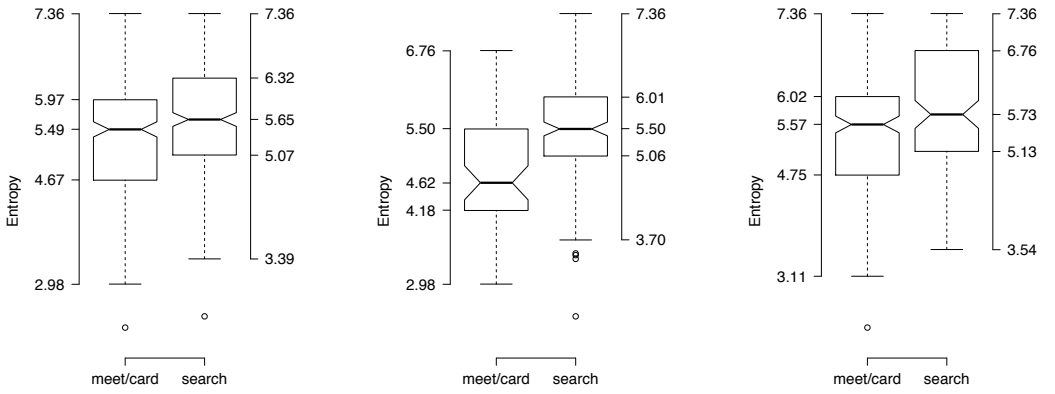


Figure 4: Smoothed denotations with associated entropy values (H), given as heatmaps in which darker colors correspond to areas of higher probability. High entropy denotations are very dispersed, whereas low entropy denotations are more concentrated. The entropy values seem to correlate well with an intuitive notion of specificity or restrictiveness.

The tasks mentioned in these hypotheses are given by the values for `engagesGoal=` in the annotations (sec. 4)). The relevant notion of specificity is given by the entropy of the probabilistic denotations (def. 7)). We can therefore rephrase the hypotheses in these measurable terms:

(14) **Specificity hypothesis (experimental version)**: Answers to questions annotated as `engagesGoal=search` will tend to have higher entropy than answers to questions annotated as `engagesGoal=card` or `engagesGoal=meet`.



(a) The full annotated set. (b) Symmetric games only. (c) Asymmetric games only.

Figure 5: The distribution of entropy values across the task types, for the whole data set and split according to whether the players had the same number of moves (symmetric) or not (asymmetric).

Fig. 5(a) is a boxplot depicting the entropy values for responses in the `engagesGoal=search` and pooled `engagesGoal=card/meet` conditions. These figures are based on 10 rounds of smoothing before the entropy calculation. The dark lines mark the median values, and the triangular indentations roughly delineate 95% confidence intervals for the medians. These intervals do not overlap, providing initial evidence that hypothesis (14) is correct. A simple linear regression using a single variable Search (= 1 if the task is `search`, 0 if it is `meet` or `card`) to predict Entropy further supports this conclusion. The coefficient for Search is estimated as 0.27 with a standard error of 0.08 ($p < 0.001$). The positive sign of this coefficient is an indication that when Search = 1, the entropy is increased as compared with Search = 0 cases, just as (14) predicts. This finding is stable across different smoothing factors; if we do no smoothing at all, the coefficient is 0.34 with a standard error of 0.17 ($p = 0.04$), and the effects get consistently stronger as we move to 10 rounds of smoothing.[2,3]

The strategies and norms of the symmetric condition differ in numerous ways from those of the asymmetric strategy, so it is worth checking to see how this distinction interacts with hypotheses (14). The boxplots in fig. 5(b) and fig. 5(c) suggest that the effect is substantial: the nature of the task has a much larger effect on specificity (entropy) in the symmetric conditions that in the asymmetric ones. I believe that this traces to the usual strategies in the two cases. In symmetric play, both players search all over the map. At any moment, they could be pretty much anywhere, so levels of specificity have to be conveyed linguistically. In asymmetric play, however, only Player 2 explores in this free way. Player 1

[2] It also holds at 20 rounds of smoothing. Beyond this, I have not systematically explored the relationship between smoothing and the strength of the effect.

[3] The player-pairings are a likely source of variation that this simple modeling ignores, so we might bring it in using a hierarchical model (Gelman & Hill, 2007; Baayen et al., 2008; Jaeger, 2008). However, perhaps owing to the large number of players and transcripts, this step seems not to notably change the coefficient estimate for Search; with both the intercept and the coefficient for Search allowed to vary by player-pairing, the fixed effect estimate for Search is 0.28 (10 rounds of smoothing) and the hierarchical estimates are relatively stable. However, I conjecture that additional exploration of these models would enhance the results described here.

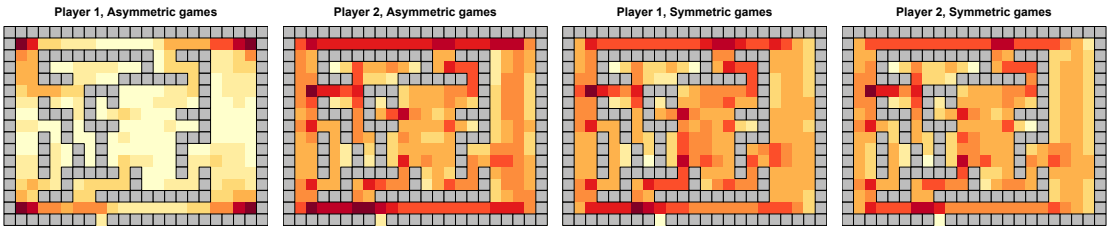


Figure 6: All the locations the players occupied, classified by player and (a)symmetry condition. In asymmetric play, Player 1 has very few moves and hence typically retreats to a corner to wait for Player 2 to bring cards (fig. 6(a)). As a result, phrases like *top right* are less vague for Player 1 in this condition.

is rendered largely immobile by her low number of available moves. The usual response to this is for Player 1 to retreat to one of the corners of the board and wait there for Player 2 to bring him cards. The corners seem to be a particularly salient and unambiguous meeting ground. The overall pattern is immediately evident in fig. 6, which depicts all of the positions that the players occupied during all of the games involving this board. The asymmetric Player 1 diagram (fig. 6(a)) stands out, with its dark corners of high activity and its hardly traveled middle area. Given this behavioral convention, Player 1 needs only to say *top right corner* to make it clear that he is going to the very top rightmost corner. This short phrase is typically high entropy for the corpus but much lower entropy in the asymmetric condition.

The clear influence of the symmetry condition on the entropy values is an indication that we should bring this source of variation in as a main effect and an interaction term. The fitted model is summarized in tab. 6. The Search coefficient is defined as before. Asymmetric is also a binary predictor. The coefficients for both are significant. Their interaction term is also significant, albeit somewhat marginally. The negative sign on the coefficient is just what one would expect from scanning the plots in fig. 5, in the sense that is shows Asymmetric conditions intuitively weakening the overall main effect of Search (even as the overall entropy is higher in asymmetric games).

	Estimate	Standard error	p value
Intercept	4.78	0.13	< 0.0001
Search	0.73	0.15	< 0.0001
Asymmetric	0.73	0.14	< 0.0001
Search*Asymmetric	-0.44	0.18	0.02

Table 6: Linear regression model highlighting the main effect of the goal-type (Search) and the influence that the Asymmetric variable has on the players' utterances.

Much more could be done to study hypothesis (14) in the corpus. The supplementary data file for this paper (available at the Cards corpus site, `http://cardscorpus.christopherpotts.net/`) contains, in addition to the information used for the above experiments, the filename for the associated transcript, the collection site, the players' Ids, the number of games each player had started at the point of utterance, the number of moves each player had available, the number of moves each player had remaining at the point of the answer, the players' current holdings, the full texts for the question and answer, and the question semantics. I'm confident that even this highly focussed and selective data set can support additional interesting experiments in goal-driven pragmatic inference.

7. Answers and the task

In this section, I introduce a modicum of decision theory in an effort to precisely explain why we find a tight correlation between the players' immediate goal and the specificity of their answers to *Where*

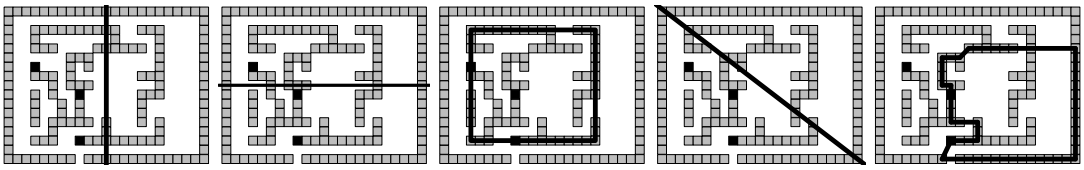


Figure 7: Some partitions for searching. Only figs. 7(a)–7(c) are easy to describe (propose), and def. 7(c) requires individual squares to be visited more than once. The boundaries are approximate. (For example, in fig. 7(b), the top player would naturally search the dead-end hallway at left.)

are you? The explanation turns on the value of new information relative to the current goal.

In a `meet` or `card` situation on gameboard $G_{m\times n}$, Player A observes a particular square $\langle i*, j*\rangle$ on $G_{m\times n}$ — either her own position (for a meet-up) or the position of a desired card. This position is unknown to Player B. The available actions for B are $a_{\langle i,j\rangle}$, interpreted as the action of B moving to square $\langle i, j\rangle$ on the board, for $0 \leqslant i < m$ and $0 \leqslant j < n$. If A utters something with semantic representation ψ, then B should go to the square with highest probability according to $[\![\psi]\!]^{G_{m\times n},s}$. In general, the lower the entropy for $[\![\psi]\!]^{G_{m\times n},s}$, the higher this probability will be. (For the data used in the experiments, entropy and maximum probability values are 96% correlated.) Thus, A should offer the lowest entropy denotation that she can, to maximize the chances that the maximum probability point is $\langle i*, j*\rangle$. Ideally, she would name a specific square in all cases. Players often fall short of this ideal, presumably for a variety of reasons — e.g., they are relying on pragmatic inferences that I have not detected in my modeling, the line-of-sight buffer permits some vagueness, the descriptions would be hard to give or hard to interpret — but they generally come close to it.

The goal in a `search` condition is very different. Here, the two players would like to partition the board into two regions, one for each player, with the size of those regions proportional to the number of moves they have. (In the asymmetric condition, Player 1 is not really fit for search, whereas the symmetric conditions favor equitable divisions of labor.) The actions are thus of the form a_{X_1,X_2} where X_1 and X_2 are the sets of points on the gameboard assigned to Player 1 and Player 2, respectively. A cost is incurred for every square that goes unvisited or is visited twice, and additional costs are incurred if the ratio of the size of the two sets X_i deviates from the ratio of the two players' allowed moves.

Concentrating on symmetric conditions to keep things simple, the above cost function favors any partition of the board into two connected, equal-sized cells. Some example partitions are given in fig. 7, in which the dark black lines mark the boundaries between search regions. It is immediately apparent that some of these partitions are easy to describe (top vs. bottom, left vs. right, inner vs. outer), whereas others would require extensive description to unambiguously propose. If we go by the players' actual behavior and vocabulary (tab. 4), then only figs. 7(a)–7(c) are reasonable proposals, and fig. 7(c) is not optimal from the point of view of search itself (though it is occasionally proposed). This provides further insight into the cost function on actions a_{X_1,X_2}: the favored ones are those in figs. 7(a)–7(b), with all others too costly to consider, either because they are too hard to describe or because they are physically inefficient. (More could be said about how the cost function balances search efficiency with descriptive efficiency. For example, in fig. 7(b), it naturally goes without saying that the top player will search the deadend hallway on the left, even though it extends into the bottom half of the board. In a sense, when a player proposes the top/bottom strategy, she takes this into account, since it would be so costly for the other player to go all the way around and into this region.)

Why are general answers favored during the `search` task? After all, naming a specific point on the board would, with some extra pragmatic reasoning, suggest an optimal search action a_{X_1,X_2}. However, at least four considerations can contribute to such answers being disfavored. First, maximally specific answers of this form are costly to convey — so much so that they are not even used with total consistency in the `meet/card` tasks. Second, such answers don't immediately suggest what search strategy the player has in mind. For example, if I note that I am inside the C room, then my partner might wonder whether I have in mind a top/bottom search action for us or a left/right one. Third, even

though specific statements like *entrance at the very bottom* would seem to indicate a top/bottom division, they don't naturally make one salient, where salience is defined in terms of lexical alternatives (Rooth, 1992; Roberts, 1996; Büring, 1999). Fourth, overly specific answers might inadvertently convey that the speaker believes himself to be pursing a `meet/card` goal — that is, pragmatic pressure from that other salient class of subgoals could militate against giving specific answers in these cases. When taken together, these considerations suggest a very articulated cost-function that, while not favoring vagueness per se, does result in it being the preferred option here.

There is one more effect that is worth noting, this one somewhat surprising from the perspective of theories that predict speakers will try to do the least work possible. Tab. 7 gives the top 10 answers in the `search` condition. As expected given the above discussion, general terms are favored, especially those that naturally suggest strategies like those depicted in fig. 7(a) and fig. 7(b). However, among the most common responses are those that pick out, not halves, but rather quadrants: *top right*, *bottom left*, and so forth. This might seem at first like at least one word's worth of unnecessary effort. However, the players seems to have figured out that it ultimately saves time. If you are on the right side of the board, the chances that the other player is also on the right are roughly 50%. Thus, saying *right* alone might not lead to a clear search strategy. If you're also in the *top* right, then the chances that the other player is there as well are just 25%. What's more, if you say *top right*, then your utterance plus the other player's is likely to immediately suggest a search strategy. The extra effort up-front saves time in dialogue.

Semantics	Count	Semantics	Count
$\textbf{BOARD}(\textbf{right})$	23	$\textbf{BOARD}(\textbf{left} \wedge \textbf{top})$	10
$\textbf{BOARD}(\textbf{middle})$	21	$\textbf{BOARD}(\textbf{left})$	9
$\textbf{BOARD}(\textbf{right} \wedge \textbf{top})$	20	$\textbf{BOARD}(\textbf{bottom} \wedge \textbf{corner} \wedge \textbf{right})$	8
$\textbf{BOARD}(\textbf{bottom} \wedge \textbf{right})$	14	$\textbf{BOARD}(\textbf{left} \wedge \textbf{middle})$	8
$\textbf{BOARD}(\textbf{bottom} \wedge \textbf{left})$	11	$\textbf{BOARD}(\textbf{hall} \wedge \textbf{right})$	7

Table 7: The top 10 utterances in the `search` scenario.

8. Conclusion

The above data and experiments provide additional support for models in which the goals and plans of the discourse participants are central to core pragmatic phenomena like resolving underspecification, determining what information is relevant, and recognizing speaker intentions. In my particular case study — locative question–answer pairs in the Cards corpus — we see wide variation in the responses people give to the single question *Where are you?*, but the variation is far from random. Rather, it is closely governed by the immediate subgoal of the discourse participants. Where the goal centers around a precise location, the responses are semantically specific. Where the goal involves a general division of the space, more general responses are preferred. The bulk of my discussion was given over to making the concepts underlying this generalization precise enough to support quantitative assessment. Such assessments bolster the initial intuition about the pragmatic factors involved, and they reveal new contours and intricacies to the phenomena.

The Cards corpus is rich enough to support more and more intricate experiments in this area. Djalali et al. (2011b) take a first look at underspecified referential expressions and quantifiers, but they explore only a small percentage of the data, and their experiments are based on version 1 of the corpus, which contains no asymmetric games. Similarly, Djalali et al. (2011a) study how the players' expertise levels affect the ways in which they negotiate the hierarchy of questions suggested by the scenario. The results provide a glimpse of the complex processes of presupposition and accommodation involved. In this paper, I focussed on locative expressions, but only those involving player locations. There is also an abundance of questions and answers about the locations of cards. The tasks that stimulate these questions are somewhat different from those studied above, and thus these interactions will provide new opportunities to study the goal-oriented nature of pragmatic inference. In sum, we essentially have a lot of promising pilots studies awaiting further research.

There are also, in the data, discourse particles, modals, a bevy of non-locative polar and constituent

questions, imperatives, emoticons, expressives, and numerous markers of certainty and uncertainty. Each bit of language can be situated in its immediate discourse context and studied from the perspective of the high-level goals of the players. I'm about to click TASK_COMPLETE on this paper, but I hope that I've made a persuasive case that others should download the corpus and use it in their own pragmatic investigations.

References

Allen, James F. (1991). *Reasoning About Plans*. Morgan Kaufmann, San Francisco.

Allen, James F., Bradford W. Miller, Eric K. Ringger & Teresa Sikorski (1996). A robust system for natural spoken dialogue. *Proceedings of the 1996 Annual Meeting of the Association for Computational Linguistics*, ACL, 62–70.

Baayen, R. Harald., Douglas J. Davidson & Douglas M. Bates (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59:4, 390–412.

Beaver, David (2002). Pragmatics, and that's an order. Barker-Plummer, David, David Beaver, Johan van Benthem & Patrick Scotto di Luzio (eds.), *Logic, Language, and Visual Information*, CSLI, Stanford, CA, 192–215.

Beaver, David & Brady Zack Clark (2008). *Sense and Sensitivity: How Focus Determines Meaning*. Wiley-Blackwell, Oxford.

Benz, Anton (2005). Utility and relevance of answers. Benz, Anton, Gerhard Jäger & Robert van Rooij (eds.), *Game Theory and Pragmatics*, Palgrave McMillan, Basingstoke, Hampshire, 195–219.

Blaylock, Nate & James F. Allen (2005). Generating artificial corpora for plan recognition. Ardissono, Liliana, Paul Brna & Antonija Mitrovic (eds.), *User Modeling 2005*, Lecture Notes in Artificial Intelligence, Springer, Berlin, 179–188.

Büring, Daniel (1999). Topic. Bosch, Peter & Rob van der Sandt (eds.), *Focus — Linguistic, Cognitive, and Computational Perspectives*, Cambridge University Press, Cambridge, 142–165.

Büring, Daniel (2003). On D-trees, beans, and B-accents. *Linguistics and Philosophy* 26:5, 511–545.

Clark, Herbert H. (1979). Responding to indirect speech acts. *Cognitive Psychology* 11:4, 430–477.

Clark, Herbert H. (1996). *Using Language*. Cambridge University Press, Cambridge.

Clark, Herbert H. & Michael F. Schober (1992). Asking questions and influencing answers. Tanur, Judith (ed.), *Questions about Questions*, Russell Sage Foundation, New York, 15–48.

Clark, Herbert H. & Deanna Wilkes-Gibbs (1986). Referring as a collaborative process. *Cognition* 22:1, 1–39.

Cover, Thomas M. & Joy A. Thomas (1991). *Elements of Information Theory*. Wiley, New York.

DeVault, David & Matthew Stone (2007). Managing ambiguities across utterances in dialogue. Artstein, Ron & Laure Vieu (eds.), *Proceedings of DECALOG 2007: Workshop on the Semantics and Pragmatics of Dialogue*.

Djalali, Alex, David Clausen, Sven Lauer, Karl Schultz & Christopher Potts (2011a). Modeling expert effects and common ground using Questions Under Discussion. *Proceedings of the AAAI Workshop on Building Representations of Common Ground with Intelligent Agents*, Association for the Advancement of Artificial Intelligence, Washington, DC.

Djalali, Alex, Sven Lauer & Christopher Potts (2011b). Corpus evidence for preference-driven interpretation. *Pre-Proceedings of the 18th Amsterdam Colloquium*, University of Amsterdam, Amsterdam.

Finucane, Cameron, Gangyuan Jing & Hadas Kress-Gazit (2010). LTLMoP: Experimenting with language, temporal logic and robot control. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1988–1993.

Frank, Michael C. & Noah D. Goodman (2012). Predicting pragmatic reasoning in language games. *Science* 336:6084, p. 998.

Franke, Michael (2009). *Signal to Act: Game Theory in Pragmatics*. ILLC Dissertation Series, Institute for Logic, Language and Computation, University of Amsterdam.

Gelman, Andrew & Jennifer Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gibbs, Raymond W. & Gregory A. Bryant (2007). Striving for optimal relevance when answering questions. *Cognition* 106:1, 345–369.

Ginzburg, Jonathan (1995a). Resolving questions, part I. *Linguistics and Philosophy* 18:5, 549–527.

Ginzburg, Jonathan (1995b). Resolving questions, part II. *Linguistics and Philosophy* 18:6, 567–609.

Ginzburg, Jonathan (1996). Dynamics and the semantics of dialogue. Seligman, Jerry (ed.), *Language, Logic, and Computation*, CSLI, Stanford, CA, vol. 1, 221–237.

Golland, Dave, Percy Liang & Dan Klein (2010). A game-theoretic approach to generating spatial descriptions. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, ACL, Cambridge, MA, 410–419.

Groenendijk, Jeroen (1999). The logic of interrogation. Matthews, Tanya & Devon Strolovitch (eds.), *Proceedings of SALT IX*, Cornell University, Ithaca, NY, 109–126.

Groenendijk, Jeroen & Floris Roelofsen (2009). Inquisitive semantics and pragmatics. Paper presented at the Stanford workshop on Language, Communication, and Rational Agency.

Groenendijk, Jeroen & Martin Stokhof (1984). *Studies in the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.

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

Jaeger, T. Florian (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59:4, 434–446.

Jäger, Gerhard (To appear). Game theory in semantics and pragmatics. Maienborn, Claudia, Klaus von Heusinger & Paul Portner (eds.), *Semantics: An International Handbook of Natural Language Meaning*, Mouton de Gruyter, Berlin.

Kress-Gazit, Hadas & George J. Pappas (2010). Automatic synthesis of robot controllers for tasks with locative prepositions. *2010 IEEE International Conference on Robotics and Automation*, IEEE, Anchorage, AK, 3215–3220.

Merin, Arthur (1997). If all our arguments had to be conclusive, there would be few of them. Arbeitspapiere SFB 340 101, University of Stuttgart, Stuttgart, URL `http://semanticsarchive.net/Archive/jVkZDI3M/`.

Merin, Arthur (1999). Information, relevance, and social decisionmaking: Some principles and results of decision-theoretic semantics. Moss, Lawrence S., Jonathan Ginzburg & Maarten de Rijke (eds.), *Logic, Language, and Information*, CSLI, Stanford, CA, vol. 2.

Parikh, Prashant (2000). Communication, meaning, and interpretation. *Linguistics and Philosophy* 23:2, 185–212.

Parikh, Prashant (2001). *The Use of Language*. CSLI, Stanford, CA.

Perrault, C. Raymond & James F. Allen (1980). A plan-based analysis of indirect speech acts. *American Journal of Computational Linguistics* 6:3–4, 167–182.

Roberts, Craige (1996). Information structure: Towards an integrated formal theory of pragmatics. Yoon, Jae Hak & Andreas Kathol (eds.), *OSU Working Papers in Linguistics*, The Ohio State University Department of Linguistics, Columbus, OH, vol. 49: Papers in Semantics, 91–136. Revised 1998.

Roberts, Craige (2004). Context in dynamic interpretation. Horn, Laurence R. & Gregory Ward (eds.), *Handbook of Pragmatics*, Blackwell, Oxford, 197–220.

Rooth, Mats (1992). A theory of focus interpretation. *Natural Language Semantics* 1:1, 75–116.

van Rooy, Robert (2003). Questioning to resolve decision problems. *Linguistics and Philosophy* 26:6, 727–763.

Stoia, Laura, Darla Magdalene Shockley, Donna K. Byron & Eric Fosler-Lussier (2008). SCARE: A situated corpus with annotated referring expressions. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, European Language Resources Association, Marrakesh, Morroco.

Stone, Matthew, Richmond Thomason & David DeVault (2007). Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. To appear in Donna K. Byron; Craige Roberts; and Scott Schwenter, *Presupposition Accommodation*.

Tellex, Stefanie (2010). *Natural Language and Spatial Reasoning*. Ph.D. thesis, MIT, Cambridge, MA.

Tellex, Stefanie & Deb Roy (2009). Grounding spatial prepositions for video search. *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, ACM, New York, 253–260.

Thompson, Henry S., Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands & Cathy Sotillo (1993). The HCRC map task corpus: Natural dialogue for speech recognition. *HLT '93: Proceedings of the workshop on Human Language Technology*, ACL, Morristown, NJ, 25–30.