Corpus evidence for preference-driven interpretation*

Alex Djalali, Sven Lauer, and Christopher Potts

Stanford University

Abstract. We present the Cards corpus of task-oriented dialogues and show how it can inform study of the ways in which discourse is goal- and preference-driven. We report on three experimental studies involving underspecified referential expressions and quantifier domain restriction.

1 Introduction

There is growing interest in the notion that both production and interpretation are shaped by the goals and preferences of the discourse participants. This is a guiding idea behind the question-driven models of Ginzburg (1996), Roberts (1996), Groenendijk (1999), and Büring (2003), as well as the related broadly decision-theoretic approaches of van Rooy (2003), Malamud (2006), Dekker (2007), and Davis (2011). These models, which we henceforth refer to generically as *goal-driven discourse models*, are intuitively well-motivated, but they have so far been tested against only a limited number of mostly hand-crafted examples and highly specific phenomena (Schoubye 2009; Toosarvandani 2010), with relatively little quantitative or corpus evaluation that we know of (but see Cooper and Larsson 2001; Ginzburg and Fernandez 2010).

The central goal of this short paper is to introduce a new publicly-available resource, the Cards corpus, and show how it can be used to explore and evaluate goal-driven discourse models (see also Djalali et al. 2011). The Cards corpus is built from a two-person online video game in which players collaboratively refine a general task description and then complete that task together. The game engine records everything that happens during play, making it possible to study precise connections between the players' utterances, the context, and their general strategies. Because the corpus is large (744 transcripts, ≈23,500 utterances) and its domain simple, it can be used to quantitatively evaluate specific pragmatic theories.

Here, we focus on the ways in which the players' conception of their task drives their understanding of referential and quantified noun phrases. After describing the corpus in more detail, we develop a simple goal-driven model that encodes and tracks certain important aspects of the players' preferences, and then we present three experiments designed to show how this model can be used to accurately resolve underspecified definites and quantifiers.

* We are indebted to Karl Schultz for designing the game engine underlying the Cards corpus. This research was supported in part by ONR grant No. N00014-10-1-0109 and ARO grant No. W911NF-07-1-0216.

2

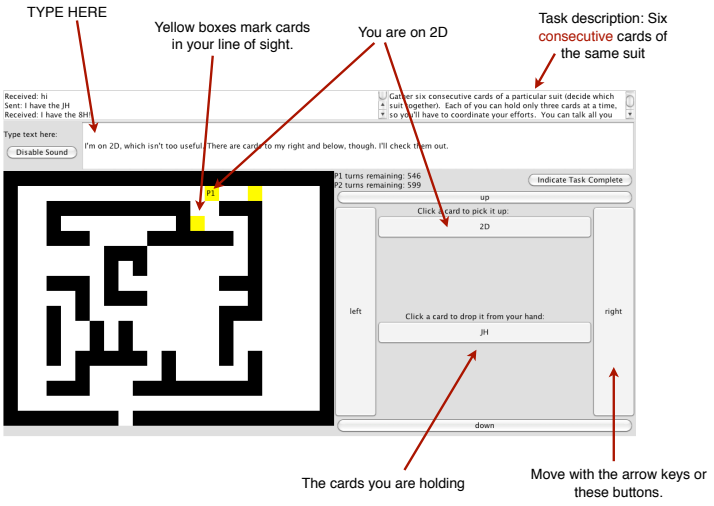

Fig. 1. An annotated version of the Cards gameboard.

2 The Cards corpus

The Cards corpus is built around a Web-based, two-person collaborative search task, partly inspired by similar efforts (Thompson et al. 1993; Allen et al. 1996; Stoia et al. 2008). We recruited players using Amazon's Mechanical Turk. The game-world consists of a maze-like environment in which a deck of cards has been randomly distributed. The players are placed in random initial positions and explore using keyboard input. A chat window allows them to exchange information and make decisions together. Each player can see his own location, but not the location of the other player. The visibility of locations of the cards are limited by distance and line-of-sight. Players can pick up and drop cards, but they can hold at most three cards at a time. In addition, while most of the walls are visible, some appear to a player only when within that player's line-of-sight.

When participants enter the game, they are presented with a description, some guidelines, and the annotated gameboard in figure 1. Before starting play, they are given the following task description (which remains visible in the upper-right of the gameboard):

> 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 general, they begin by wandering around reporting on what they see, exchanging information as they go

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	MAX_LINEOFSIGHT	3
Server	2	MAX_CARDS	3
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

Table 1. Environment metadata in the corpus format.

until a viable strategy begins to emerge. Dialogue (1) is typical. (Between utterances, the players explore the environment and manipulate cards; this dialogue spans a total of 56 moves.)

(1) P1: i am top right
P2: im bottom left
P1: ok
P2: i have 3, 7h
P2: also found 6 and 7s...what should we go with?
P1: moving down to the bottom on that long corridor not seen a heart yet
P1: you pick
P2: u have the 9s right?
P1: yep
P2: ok lets go spades i have the 7 and 6

And then they pursue a specific solution. Because they can hold only three cards at a time, they are compelled to share information about the locations of cards, and their solutions are necessarily collaborative.

The current release (version 1) consists of 744 transcripts. 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 23,532 utterances (mean length: 5.84 words), totaling 137,323 words, with a vocabulary size around 3,500. Most actions are not utterances, though: there are 255,734 movements, 11,027 card pick-ups, and 7,202 card drops. The median game-length is 414 actions, though this is extremely variable (standard deviation: 261 actions).

The transcripts are in CSV format. Table 1 is an example of the high-level environmental information included in the files, and table 2 is a snippet of gameplay. Computationally, one can update the initial state to reflect the players' actions, thereby deriving from each transcript a sequence of ⟨context, event⟩ pairs. This makes it easy to study players' movements, to make inferences about what

Agent	Time	Action type	Contents
Player 1	566650	PLAYER_MOVE	7,11
Player 2	567771	CHAT_MESSAGE_PREFIX	which c's do you have again?
Player 1	576500	CHAT_MESSAGE_PREFIX	i have a 5c and an 8c
Player 2	577907	CHAT_MESSAGE_PREFIX	i jsut found a 4 of clubs
Player 1	581474	PLAYER_PICKUP_CARD	7,11:8C
Player 1	586098	PLAYER_MOVE	7,10

Table 2. A snippet of gameplay in the corpus format.

guides their decision-making and, most importantly for pragmatics, to study their language in context.

The Cards corpus is available at http://cardscorpus.christopherpotts.net/. The distribution includes the transcripts, starter code for working with them in Python and R, and a slideshow containing documentation. We think that the corpus fills an important niche; while there are a number of excellent task-oriented corpora available, the Cards corpus stands out for being large enough to support quantitative work and structured enough to permit researchers to isolate very specific phenomena and make confident inferences about the participants' intentions.

3 Relevance and the evolving task

The strategic aspects of interactions in the Cards corpus revolve around sequences of cards. In Djalali et al. 2011, we developed a hierarchy of abstract questions concerning the game, cards, and strategies, and showed that expert players and novice players negotiate this hierarchy differently: novices work systematically through it, whereas experts strategically presuppose resolutions of general issues so that they can immediately engage low-level task-oriented ones.

Here, we focus on how the relevance of particular cards changes as the players' strategies change. To this end, we define the *value* of a hand H, $\mathit{Value}(H)$, to be the minimum number of pick-up and drop moves from H to a solution to the game. For example, $\mathit{Value}(\emptyset) = 6$, $\mathit{Value}(\{5\text{H}\}) = 5$, and $\mathit{Value}(\{5\text{H}, 8\text{H}\}) = 4$. Because dropping is costly, $\mathit{Value}(\{5\text{H}, 2\text{S}, 3\text{D}\}) = 7$: at least two cards have to be dropped before forward progress towards a solution.

The *Value* function gives rise to a measure of how relevant a card c is given a hand H:

$$\mathit{Relevance}(c, H) \stackrel{def}{=} V(H) - V(H \cup \{c\}) \tag{2}$$

Where c is intuitively relevant to H, this value is $+1$, else it is -1 (or 0 if $c \in H$). Figure 2 illustrates the relevance sphere for a particular hand $H = \{2\text{H}, 4\text{H}, 5\text{H}\}$.

Our overarching hypothesis is that the players will seek cards that are relevant given their current holdings, and that this will be a driving force in how they resolve linguistic underspecification. The experiments in the next section seek

to refine and support this basic idea. We should say, though, that the current notion of relevance is just an approximation of the players' underlying policy. We know, for example, that they often pick up irrelevant cards that might be relevant later — the cost of refinding cards is greater than the cost of having to drop those that turn out to be irrelevant. Our measure also does not take into account the costs of communication: the players have a slight bias for hearts, probably because this is the most iconic of the suits; they are reluctant to change suits once they have settled on one (even if this means extra exploration); and they favor solutions that don't span the Ace, since there is indeterminacy about whether such solutions are legitimate. We are confident that our results will only get stronger once these considerations are brought into the modeling. We mention them largely to emphasize that the corpus can support sophisticated investigations into decision making and pragmatics together.

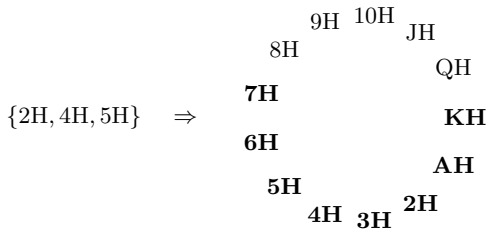

Fig. 2. Given the hand at left, the cards in bold are relevant in the sense that obtaining any one of them would move the players closer to a solution. The others would take the players farther from a solution, as measured by pick-up and drop moves, and are thus irrelevant given this hand.

4 Experiments

4.1 Experiment 1: underspecified card references

It is a testament to the importance of context that the majority of singular card references in the corpus (see table 3(a)) are like those in the following dialogue:

(3) P2: Look for 2.
P1: and the 3?

Absent knowledge of the context, we cannot know which cards these players are referring to. However, if we know that they are collectively holding 4H and KH, then their intentions become clear: P2 refers to 2H and P1 to 3H. Intuitively, such resolutions maximize relevance. More specifically, we hypothesized that, for any nominal referring expression N and stage of play t, the intended referent will be in the set

(4) $\text{argmax}_{t,c \in Res(N)}\ Value(H_t) - Value(H_t \cup \{c\})$

Phrase type	Count
Fully specified	103 (37%)
Underspecified	172 (63%)
Total	275

(a) Singular definite card references

Inference	Count
Correct	164 (95%)
Incorrect	8 (5%)
Total	172

(b) Results.

Table 3. Resolving underspecified card references via relevance.

where H_t is the set of cards the players are currently holding at t, and $Res(N)$ is the set of cards consistent with the descriptive content of N. (For example, $Res(4) = \{4\text{H}, 4\text{D}, 4\text{S}, 4\text{C}\}$ and $Res(H) =$ the set of all hearts.)

To evaluate this hypothesis, we annotated the underspecified singular card references in 10 transcripts for what we took to be the intended referents and then wrote a computer program that chose the maximally relevant interpretation according to (4). (Where there was more than one such maximum, the program chose one at random.) We count a prediction as correct iff it matches the human annotation. The accuracy of this algorithm is extremely high (table 3(b)). What's more, we find that its mistakes tend to be clustered together near the start of transcripts, where even the interpreting player might have felt unsure about the speaker's intentions.

4.2 Experiment 2: unrestricted quantification?

From a strict logical perspective, quantifiers like *everything*, *nothing*, and *anything* carry universal force: the truth of a sentence involving them requires checking that every entity in the domain of discourse has (or does not have) the property they apply to. In this sense, P1 speaks falsely in (5) when he says, "I see nothing". We know this because the rich meta-data of the Cards corpus allows us to calculate exactly which cards P1 saw prior to this utterance. (He happened to have seen 5C and 10S.)

(5) P1: ok–i'll look at D and H, u look at C and S?
P2: ok
P1: i see nothing

It is clear why P1 is not perceived as speaking falsely, though: at the time of utterance, he had seen no cards that were relevant to his initial proposal, that is, no diamonds or hearts. In (6), the implicit restrictions are even more refined:

(6) P1: lets do spades
P1: I have the as, qs, and ks
[. . .]
P1: ok, i found js

Quantifier	Literally true	Literally false
anything	2	6
nothing	0	6
Total	2 (14%)	12 (86%)

Table 4. Experiment 2 results.

P2: Ok. I haven't found anything...lol

Here, P2's *anything* seems to range just over the cards that are relevant to the hand $\{\text{AS}, \text{QS}, \text{KS}\}$, in the sense of (2). That is, P2 saw other cards, just not in this contextually privileged set.

The highly constrained nature of the Cards world means that we can precisely define the domain of discourse, which in turn permits us to identify exactly which contextual factors shape the domain in these cases. The first question we sought ask in this area is a seemingly simple one: what percentage of universally quantified claims are literally true, that is, true for an unrestricted interpretation of their quantifiers? To make our experiments manageable, we first extracted all utterances matching the regular expression in (7):

(7) `(find|found|see|saw) (any|no)thing`

Such phrases include simple declaratives like *I see nothing* as well as interrogatives like *Did you find anything?* Although there are 35 matches for (7), we disregard ones like *I didn't see anything around here*, as they overtly restrict the domain of discourse and often involve indexical terms, whose semantics are more difficult to define.

We define one of these quantified phrases as *literally true* just in case it is true on an unrestricted interpretation of the quantifier, that is, a quantifier that ranges over the full deck of 52 cards. At any point t in the game, each player P will have walked over, and hence seen/found ('seeing' and 'finding' amount to the same thing in the Cards world, as players can identify only the cards they are currently standing on), a subset of the full deck of cards. Call this subset $S_{P,t}$. Suppose that, at time t, player P says, "I found nothing". Then player P's claim is literally true just in case $S_{P,t} = \emptyset$ and literally false just in case $S_{P,t} \neq \emptyset$. If, at time t, a player P asks "Did you find anything" and player P' responds "Yes", then P''s response is literally true just in case $S_{P,t} \neq \emptyset$, and literally false just in case $S_{P,t} = \emptyset$. (Similarly if P' responds with 'No'.)

These definitions mean that we can quantitatively assess the percentage of universal claims that are literally true. The procedure is as follows: for each universally quantified claim made by a player P at time t, build $S_{P,t}$ by following the full path taken by P up to t and adding to $S_{P,t}$ all and only those cards walked over by P up to t. Table 4 summarizes the results of this experiment. As is evident, effectively no quantified phrases are literally true. Indeed, the only time players speak literally is during the initial stages of the game when they

are trying to establish what sequence they should pursue, as in (8), in which P2 has found and is holding only the QC.

(8) P1: ok so what suit
P2: Whatever I find first.
[. . .]
P1: have you found anything yet
P2: I have QC.

4.3 Experiment 3: goal-based domain restriction

The results of experiment 2 indicate that implicit quantifier restrictions are the norm. We turn now to the task of identifying which contextual factors the players use to provide these restrictions. It turns out the players' decisions about which suit to pursue are reliable indicators of how the domain of discourse is restricted. In the following dialogue, the players agree to pursue clubs:

(9) P1: lets go clubs
[. . .]
P2: ok I finished right side and middle box did you find anything?
P1: no

Literally speaking, P1 has spoken falsely, as (s)he found 11 cards prior to uttering "no". However, none of those findings include clubs. Thus, P1 speaks truthfully if we regard the earlier negotiation as restricting the domain for *anything* in P2's utterance.

In the terms of section 3, it looks like P1's proposal to limit to clubs makes all and only those cards relevant, which provides the basis for interpreting the quantifier. To test this idea, we hand-annotated all the transcripts involving the stimuli in experiment 2 for the players' mutually agreed-upon suit. The annotations mark the span of text in which the negotiation occurred with the suit they agreed upon. Phrases that indicated the start of such a negotiation included, but were not limited to, "let's go (for) X", "look for X", and simple suit mentions like "hearts?".

We now define one of the quantified phrases in (7) as *restrictedly true* just in case it is true on a restricted interpretation of the quantifier, that is to say, a quantifier that ranges only over cards with the agreed-upon suit. (A prominent edge case involved players who never overtly agreed upon any suit in particular, but rather deployed a strategy of stacking cards in a large pile and looking for any winning sequence. In such cases, we made the simplifying assumption that the domain of discourse included all and only the winning sequences.)

We were able to annotate the transcripts of 12 of the 14 quantified phrases considered in experiment 1. We disregarded the two phrases that involved literally true uses of the quantifiers, as they were used when players are trying to establish what sequence they should agree upon, and reran the same experiment as above. The results are given in table 5. They are essentially the opposite of

Quantifier	Literally true	Literally false
anything	5	1
nothing	6	0
Total	11 (92%)	1 (8%)

Table 5. Experiment 3 results.

those in table 4: when we restrict the domain of discourse to the players' agreed upon suit, utterances that were false become true. This is precisely the result one would expect on a model of discourse in which interpretation is governed by high-level factors relating to the discourse participants' understanding of the goals and issues at hand.

5 Conclusion

This paper introduced the Cards corpus, a highly-structured resource for doing corpus-driven computational pragmatics. In a series of experiments, we showed how the transcripts can be used to precisely define the domain, to ground denotations for quantified terms, and to pinpoint ways in which the context influences utterance understanding.

These experiments show that the Cards corpus can support quantitative evaluation of hypotheses in pragmatics. Going forward, we hope to expand our theoretical reach by pursuing the following inter-related goals:

- Increase the size and power of the corpus by collecting additional transcripts and altering the parameters of the game, e.g., number of moves, line of sight, and number of cards each player is able to hold.
- Extend our experimental techniques to a wider range of phenomena already present in the corpus. For example, there are many utterances in the corpus of the form "4H" or even just "4". In context, it is clear what the speaker intends: "Found the 4H", "Can't find the 4H", "Look for the 4H", etc. We conjecture that, just as the context can be used effectively to resolve the underspecification of "4" (section 4.1), so too can it be used to resolve which predicate the speaker intended.
- Situate the above results in a fuller question-driven model of the sort employed by Djalali et al. 2011. Both implicit and explicit questions shape players' actions (where to move, what to pick up, when to speak, and so forth). For example, the players' negotiations about which suit to pursue (experiment 3) fit neatly into a goal-driven question hierarchy of the sort envisioned by Roberts (1996). Using a question model, we can study the players' linguistic behavior, and we can pursue questions in pragmatics and decision theory simultaneously, finding new ways in which language shapes, and is shaped by, the goals and preferences of the discourse participants.

Bibliography

Allen, James F.; Bradford W. Miller; Eric K. Ringger; and Teresa Sikorski. 1996. A robust system for natural spoken dialogue. In *Proceedings of ACL*, 62–70.

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

Cooper, Robin and Staffan Larsson. 2001. Accommodation and reaccommodation in dialogue. In *Göteborg Papers in Computational Linguistics*. Department of Linguistics, Göteborg University.

Davis, Christopher. 2011. *Constraining Interpretation: Sentence Final Particles in Japanese*. Ph.D. thesis, UMass Amherst.

Dekker, Paul. 2007. Optimal inquisitive discourse. In Maria Aloni; Alistair Butler; and Paul Dekker, eds., *Questions in Dynamic Semantics*, 83–101. Elsevier.

Djalali, Alex; David Clausen; Sven Lauer; Karl Schultz; and Christopher Potts. 2011. Modeling expert effects and common ground using Questions Under Discussion. In *Proceedings of the AAAI Workshop on Building Representations of Common Ground with Intelligent Agents*. AAAI.

Ginzburg, Jonathan. 1996. Dynamics and the semantics of dialogue. In Jerry Seligman, ed., *Language, Logic, and Computation*. CSLI.

Ginzburg, Jonathan and Raquel Fernandez. 2010. Computational models of dialogue. In Alex Clark; Chris Fox; and Shalom Lappin, eds., *Handbook of Computational Linguistics and Natural Language Processing*. Blackwell.

Groenendijk, Jeroen. 1999. The logic of interrogation. In Tanya Matthews and Devon Strolovitch, eds., *Proceedings of SALT IX*, 109–126. Cornell University.

Malamud, Sophia. 2006. *Semantics and Pragmatics of Arbitrariness*. Ph.D. thesis, Penn.

Roberts, Craige. 1996. Information structure: Towards an integrated formal theory of pragmatics. In Jae Hak Yoon and Andreas Kathol, eds., *OSU Working Papers in Linguistics*, 91–136. The Ohio State University Department of Linguistics. Revised 1998.

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

Schoubye, Anders. 2009. Descriptions, truth value intuitions, and questions. *Linguistics and Philosophy* 32(6):583–617.

Stoia, Laura; Darla Magdalene Shockley; Donna K. Byron; and Eric Fosler-Lussier. 2008. SCARE: A situated corpus with annotated referring expressions. In *Proceedings of LREC*.

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

Toosarvandani, Maziar. 2010. *Association with Foci*. Ph.D. thesis, UC Berkeley.