Computational challenges in explaining communication: How deep the rabbit hole goes

Laura van de Braak (l.vandebraak@donders.ru.nl)
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Mark Dingemanse (m.dingemanse@let.ru.nl)
Centre for Language Studies, Radboud University, The Netherlands

Ivan Toni (ivan.toni@donders.ru.nl)
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Iris van Rooij (i.vanrooij@donders.ru.nl)
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Mark Blokpoel (m.blokpoel@donders.ru.nl)
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Abstract

When people are unsure of the intended meaning of a word, they often ask for clarification. One way of doing so—often assumed in models of communication—is to point at a potential target: "Do you mean [points at the rabbit]?" However, what if the target is unavailable? Then the only recourse is language itself, which seems equivalent to pulling oneself up from a swamp by one's hair. We created two computational models of communication, one able to point and one not. The latter incorporates inference to resolve the meaning of non-pointing signals. Simulations show agents in both models reach perceived understanding equally quickly. While this means agents think they are successfully communicating, non-pointing agents understand each other only at chance level. This shows that state-of-the-art computational explanations have difficulty explaining how people solve the puzzle of underdetermination, and that doing so will require a fundamental leap forward.

Keywords: communication; pragmatics; Bayesian modeling; agent-based modeling; computational modeling

Introduction

The puzzle of language use in interaction is elegantly illustrated by Quine's famous *gavagai* thought experiment (Quine, 1960): A rabbit scuttles across and someone shouts, "gavagai!" How do you understand what the word means? The difficulty of this puzzle is illustrated by the fact that 60 years later, we are still only scratching the surface. We know people can understand each other, it is just very hard to explain how. Now consider the following variant to discover how deep this rabbit hole goes: The other shouts "gavagai!" You may wish to point at the rabbit asking "gavagai?", but alas it has run away. How can you still make sense of the word now that you are unable to point at your displaced reference? Your only recourse is language itself, but now that you are in a swamp of underdetermination how can you pull yourself up by your own hair?

Prior solutions to the puzzle have leaned heavily on *repair actions*. Communicators can shore up understanding piecemeal by asking and providing clarification (Dingemanse et al., 2015; Schegleff, Jefferson, & Sacks, 1977; Clark & Schaefer, 1987). While people seem flexible enough to deal with a great deal of referential inscrutability (Garfinkel, 1967), computational models of communication tend to assume more immediate forms of feedback: for instance by ostension (pointing at an intended or inferred referent) or even providing direct shared access (Steels, 2015; Hawkins, Frank, & Goodman, 2017).[1] Even under these ideal conditions, it has proven difficult to computationally explain how two communicators can come to understand one another. State-of-the-art models operate on small contexts and languages, and they require uncharacteristically many trial-and-error attempts or fail to reach mutual understanding (Steels, 2015). We do not take this merely as a negative result, but as an indication of a truly hard to explain aspect of human communication. It is challenging to explain how people can pull themselves out of the swamp with an overhanging branch (direct access), let alone by their own hair (ostensive signals), let alone when they pull and realise they are wearing a wig (non-ostensive signals).

In this paper, we present and compare two computational models of communication. One to model cases where people have access to unambiguous ostensive pointing signals, and one to model cases where people rely on additional inference to resolve the meaning of non-ostensive signals. Agent-based simulations show that having access to ostensive signals allows agents to converge on a common lexicon and achieve relatively successful referential communication. The agents that rely on inferences over non-ostensive signals enjoy much less success, despite their ability to leverage additional inference. This shows that sophisticated inferential abilities are not enough to overcome the referential inscrutability of non-ostensive signals.

To understand why, we distinguish between factual and perceived states of understanding (Fig. 1). In doing so, we contribute to prior work questioning dogmas of understand-

[1] The term ostensive has two senses: (a) referential transparency, i.e., no (or little) reasoning is required to understand the meaning of the signal, or (b) a signal that communicates its own communicative value (Wilson & Sperber, 2002). Here, we use it in the first sense.

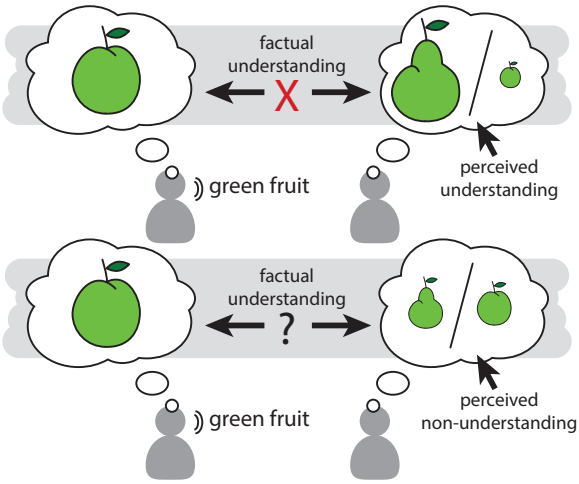


Figure 1: Two kinds of understanding. Without direct access to other minds, people can only use their own *perceived understanding* as a proxy for *factual understanding*. When they are in a state of perceived understanding, they assume there is also factual understanding, which may or may not be correct (top); and when they are in a state of perceived non-understanding, i.e., they are uncertain about their understanding, they may not commit to any interpretation in which case factual understanding remains undetermined (bottom).

ing in theories of human interaction (Clark, 1997). The first refers to the state where both agents have the same referent in mind (Clark, 1996), e.g., both think that gavagai refers to the rabbit's nose. Note that this state is a property defined over two communicators, and that people cannot know if this state is true or not. By contrast, perceived understanding is an individual state. Namely, whether or not a person 'thinks' they have correctly understood the communicative signal. In a follow-up analysis we show that, differently from the ostensive agents, non-ostensive agents incorrectly think that they made correct inferences and decide to end the dialogue in a state of factual misunderstanding. In our discussion we reflect on theoretical challenges and opportunities for the field.

Computational models

We present two computational models based on the Rational Speech Act (RSA) framework (Frank & Goodman, 2012). RSA formalizes Gricean cooperative communication (Grice, 1975) as Bayesian rational communication (Frank, Emilsson, Peloquin, Goodman, & Potts, 2017). Speakers choose signals relative to the probability that the chosen signal will be interpreted correctly (taking the listener's perspective), and listeners infer meaning relative to the probability that a speaker would have produced the observed signal given that meaning (taking the speaker's perspective). RSA has been successfully used to explain language games (Frank & Goodman, 2012; Frank et al., 2017; Khani, Goodman, & Liang, 2018), implicatures (Goodman & Stuhlmüller, 2013; Bergen, Levy, & Goodman, 2016), noisy signals (Bergen & Goodman, 2015), polite speech (Yoon, Frank, Tessler, & Goodman, 2018), and convention formation (Hawkins et al., 2017; Cohn-Gordon, Goodman, & Potts, 2018; Hawkins, Frank, & Goodman, 2020).

We take the convention formation model by Hawkins et al. (2017) as a starting point to model ostensive and non-ostensive dialogue. The model explains how a listener can, on the basis of previous ostensive communicative exchanges, infer the probability of the possible signal-referent mappings the speaker is using and uses this to infer the most likely intended meaning. We extend this model in two ways.

First, we extend it to capture *ostensive dialogue* between two agents where, differently from existing models, both agents can contribute to the conversation by speaking and pointing. Ostensive communicators can request clarification analagous to asking "gavagai?" and pointing at the rabbit, making referentially clear what they believe the word "gavagai" refers to. Second, we provide a further extension to capture *non-ostensive dialogue* where ostensive signals are not available. The crucial difference here is that the clarification request 'gavagai?' is not accompanied by an ostensive signal. In this model communicators are endowed with Bayesian inference to infer the meaning of the clarification signal, instead of having it unambiguously provided.

In these models both agents produce and interpret signals, so we introduce the terms *initiator* and *responder*. The initiator has the intention to communicate a particular referent and starts the dialogue. The responder is trying to best understand which referent the initiator means and can request clarification to which the initiator can reply. This process repeats, moving the dialogue forward. Dialogues in both models can end in one of three ways (see Fig. 2): either the responder believes they understood the initiator, or the initiator believes that the responder understood them, or they give up after several turns. We next introduce RSA before we formalize ostensive and non-ostensive dialogue.

Rational Speech Act model

RSA provides a characterization of pragmatic speaking and listening at the computational level (Frank & Goodman, 2012). It assumes that communicators can take each others' perspective by reasoning recursively, determined by parameter n. For example, a pragmatic listener will infer "fruit" refers to the apple, not the cherry, if the only other word avaiable is "red". RSA abstracts language to a many-to-many mapping between signals S and referents R. Such a mapping is called a *lexicon* $\mathcal{L}$. The speaking and listening capacities formalized by RSA form the basis for the initiator and responder models below. Higher order speaking is defined as:

$$S_n(s|r,\mathcal{L}) \propto exp\left(\alpha\, log\, L_{n-1}\left(r|s,\mathcal{L}\right) - cost(s)\right) \qquad (1)$$

This probability distribution over signals given a referent and lexicon depends on the (presumed) inference made by a lis-

tener of one order lower L_{n-1}:

$$L_n(r|s,\mathcal{L}) \propto Pr(r)\ S_n(s|r,\mathcal{L}) \tag{2}$$

The recursion bottoms out in a literal listener model L_0 that takes the lexicon on face value, apart from incorporating a prior probability for each referent $\Pr(r)$:

$$L_0(r|s,\mathcal{L}) \propto Pr(r)\ \mathcal{L}(s,r) \tag{3}$$

In this paper we assume uniform priors over referents.

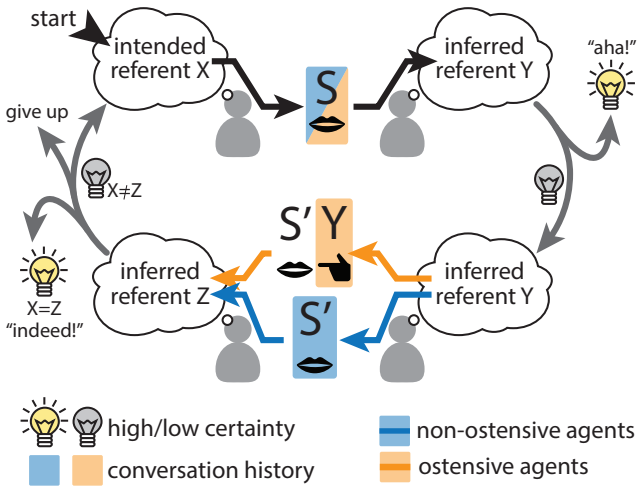


Figure 2: Structure for ostensive (orange) and non-ostensive (blue) dialogue for one intended referent X. Black, orange and blue arrows denote input-output of information. Gray arrows denote agents' decisions for continuing or ending the dialogue. Agents remember all turns and take this into account for future inferences. The key difference is that ostensive agents base future inferences on the ostensively declared referent Y (orange rectangle) whereas the non-ostensive agents base future inferences on a verbal signal S' (blue rectangle). The initiator infers referent Z from S' to determine whether they think that the responder understands them.

Formalizing ostensive communication

In ostensive communication, the responder can request clarification by producing a pointing gesture. Both agents remember all signals spoken by the intiator s_{initial} and the corresponding clarification gesture which is formalized as the inferred referent r_{inferred} (S and Y in the orange boxes in Fig. 2). Agents use conversation history h and their lexical bias $\Pr(\mathcal{L})$ to infer the likelihood over all possible lexicons (Fig. 3):

$$\Pr_{L_n}(\mathcal{L} \mid h) \propto \Pr(\mathcal{L}) \prod_{(s,r)\in h} L_n(r|s,\mathcal{L}) \tag{4}$$

The distribution over referents given a signal and history can now be formalized as:

$$\Pr_{L_n}(r \mid s,h) \propto \sum_{\mathcal{L}} L_n(r|s,\mathcal{L}) \Pr_{L_n}(\mathcal{L}|h) \tag{5}$$

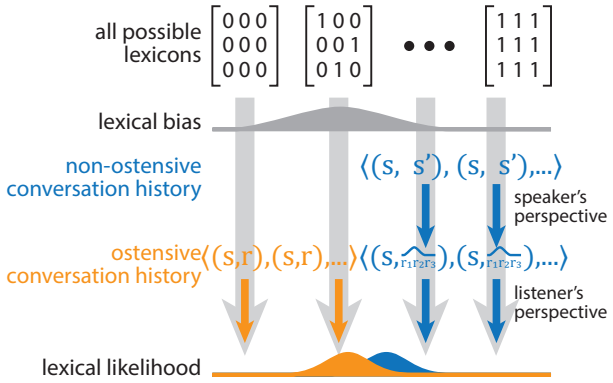


Figure 3: The distribution over all lexicons (lexicon likelihood) is computed from the lexical bias and conversation history. Non-ostensive agents (blue) require additional inference to infer the meaning of clarification requests s', whereas ostensive agents (orange) get unambiguous feedback r.

This distribution is used by both ostensive agents. The responder infers a referent given an observed signal by sampling r_{inferred} from $\Pr_{L_n}(r \mid s_{\text{observed}}, h)$. The initiator does the same but given the observed clarification request s_{request}, to try and understand the responder. Both agents use the entropy $H(.)$ of this distribution to decide if they are certain enough of their inference, formalizing *perceived understanding*. The responder makes a clarification request when uncertain, by sampling a signal s_{request} given their inferred referent r_{inferred} from:

$$\Pr_{S_n}(s \mid r,h) \propto e^{\left(\alpha \ln\left(\sum_{\mathcal{L}} \Pr_{L_n}(\mathcal{L}|h)\ L_{n-1}(r|s,\mathcal{L})\right) - cost(s)\right)} \tag{6}$$

Here, the distribution over signals takes into account history, all possible lexicons and pragmatic reasoning. The cost function could represent for example utterance length, and α is the rationality parameter (higher values make sampling the most probable signal more likely). The ostensive responder can both express s_{request} *and* point unambiguously to their inferred referent r_{inferred}, where s_{request} is sampled from $\Pr_{S_n}(s \mid r_{\text{inferred}}, h)$. The agents signal a special change-of-state token (cf. Koivisto, 2015; Selting, 1987) to indicate when they perceive understanding: The responder can use `aha!` to confirm they think they correctly inferred the intended referent r_{intended}. Similarly, the initiator can use an acknowledgement token like `indeed!` when they think the responder has correctly inferred the intended referent (Clark, 1996). Table 1 illustrates an example conversation between agents. For completeness we provide the formalization of the ostensive responder and initiator below.

OSTENSIVE RESPONDER

Input: A set of signals S, a set of referents R, and the conversation history $h = ((s_{\text{initial}}, r_{\text{inferred}}), \dots)$. An observed signal $s_{\text{observed}} \in S$ and an order of reasoning n. A linguistic bias representing the agent's preferred lexicons $\Pr(L)$. A rationality parameter α, a certainty threshold η and a cost function $cost : S \to \mathbb{N}$.

Output: A clarification request $(s_{\text{request}}, r_{\text{inferred}})$ if $H(\Pr_{L_n}(r \mid s_{\text{observed}}, h)) > \eta$; or otherwise the inferred referent r_{inferred} and a special confirmation signal `aha!` Here, s_{request} is sampled from $\Pr_{S_n}(s \mid r_{\text{inferred}}, h)$ (Eq. 6) and r_{inferred} is sampled from $\Pr_{L_n}(r \mid s_{\text{observed}}, h)$ (Eq. 5).

OSTENSIVE INITIATOR

Input: A set of signals S, a set of referents R, and the conversation history $h = ((s,r),\dots)$. A referential intention $r_{\text{intended}} \in R$ and an order of reasoning n. A linguistic bias representing the agent's preferred lexicons $\Pr(L)$. A rationality parameter α, a certainty threshold η and a cost function $cost : S \to \mathbb{N}$. Optionally, an observed clarification request $s_{\text{request}} \in S$.

Output: The initial signal s_{initial} sampled from $\Pr_{S_n}(s \mid r_{\text{intended}}, h)$ (Eq. 6) if no clarification request was made, a clarification response s_{response} sampled from the same distribution if $H(\Pr_{L_n}(r \mid s_{\text{request}}, h))$ is above η (Eq. 5), or a special confirmation signal `indeed!` if entropy is below η and r_{intended} matches r_{inferred} sampled from $\Pr_{L_n}(r \mid s_{\text{request}}, h)$ (Eq. 5).

Formalizing non-ostensive communication

In non-ostensive communication, communicators cannot use ostensive signalling and build up a conversation history consisting only of initiator signals s_{initial} and corresponding clarification requests s_{request} (S and S' in the blue boxes in Fig. 2). Agents use this conversation history h and their lexical bias $\Pr(\mathcal{L})$ to infer the likelihood over all possible lexicons. Differently from the ostensive model, non-ostensive agents make an additional inference interpreting the clarification request $(L_n\left(r|s_{\text{request}}, \mathcal{L}\right)$ below; Fig. 3):

$$\Pr_{L_n}(\mathcal{L} \mid h) \propto \Pr(\mathcal{L}) \prod_{(s_{\text{initial}}, s_{\text{request}}) \in h} \sum_{r} S_n\left(s_{\text{initial}} | r, \mathcal{L}\right) L_n\left(r | s_{\text{request}}, \mathcal{L}\right) \tag{7}$$

The distribution over referents given a signal and history $\Pr_{L_n}(r \mid s,h)$ is formalized as Eq. 5 replacing $\Pr_{L_n}(\mathcal{L} \mid h)$ with Eq. 7. Similarly to the ostensive agents, non-ostensive initiator and responder infer referential meaning of signals r_{inferred} and their perceived certainty thereof from this distribution. The responder makes a clarification request when uncertain by sampling a signal s_{request} given their inferred referent r_{inferred} from $\Pr_{S_n}(s \mid r,h)$ (Eq. 6), again replacing $\Pr_{L_n}(\mathcal{L} \mid h)$ with Eq. 7.

NON-OSTENSIVE RESPONDER

Input: Same as OSTENSIVE RESPONDER, except the conversation history $h = \left((s_{\text{initial}}, s_{\text{request}}), \dots\right)$ consists of pairs of initiator signals s_{initial} and the responder's clarification requests s_{request}.

Output: A clarification request s_{request} if $H(\Pr_{L_n}(r \mid s_{\text{observed}}, h)) > \eta$ (Eq. 5, Eq. 7); or otherwise a the inferred referent r_{inferred} and a special confirmation signal `aha!` Here, s_{request} is sampled from $\Pr_{S_n}(s \mid r_{\text{inferred}}, h)$ (Eq. 6, Eq. 7) and r_{inferred} is sampled from $\Pr_{L_n}(r \mid s_{\text{observed}}, h)$ (Eq. 5, Eq. 7).

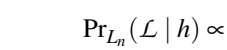
NON-OSTENSIVE INITIATOR

Input: Same as OSTENSIVE INITIATOR, except that the conversation history $h = \left((s_{\text{initial}}, s_{\text{request}}), \dots\right)$ consists of pairs of initiator signals s_{initial} and the responder's clarification requests s_{request}.

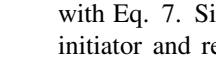
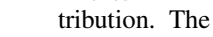
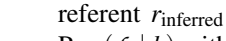
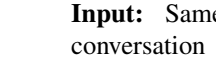
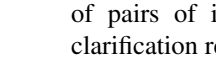
Output: The initial signal s_{initial} sampled from $\Pr_{S_n}(s \mid r_{\text{intended}}, h)$ (Eq. 6, Eq. 7) if no clarification request was made, or else a clarification response s_{response} sampled from the same distribution if $H(\Pr_{L_n}(r \mid s_{\text{request}}, h))$ is above η (Eq. 5, Eq. 7), or a special confirmation signal `indeed!` if entropy is below η and r_{intended} matches r_{inferred} sampled from $\Pr_{L_n}(r \mid s_{\text{request}}, h)$ (Eq. 5, Eq. 7).

Table 1: Example conversations between ostensive (a) and non-ostensive (b) agents for toy 2-by-2 lexicons (six possible in total). Lexical likelihoods $\Pr_{L_n}(\mathcal{L} \mid h)$ at t_i are based on lexical bias and history (indicated in bold) from $t_1 \dots t_{i-1}$. Ostensive agents have access to explicit feedback (i.e., referents pointed out by the responder), whereas non-ostensive agents need to make do with ambiguous linguistic feedback. The ostensive agents reach perceived understanding after two turns, where the non-ostensive agents give up after six turns.

(a) Conversation between ostensive agents.

	$\Pr_{L_n}(\mathcal{L} \mid h)$	Initiator		Responder	$\Pr_{L_n}(\mathcal{L} \mid h)$
t_1		$r_2 \rightarrow$ $r_1 \leftarrow$	$\mathbf{s_1}$ $(s_2, \mathbf{r_1})?$	$\rightarrow r_1$ $\leftarrow r_1$	
t_2		$r_2 \rightarrow$	$\mathbf{s_2}$ `aha!`	$\rightarrow r_2$ $\leftarrow r_2$	

(b) Conversation between non-ostensive agents.

	$\Pr_{L_n}(\mathcal{L} \mid h)$	Initiator		Responder	$\Pr_{L_n}(\mathcal{L} \mid h)$
t_1		$r_1 \rightarrow$ $r_2 \leftarrow$	$\mathbf{s_2}$ $\mathbf{s_1}?$	$\rightarrow r_2$ $\leftarrow r_2$	
t_2		$r_1 \rightarrow$ $r_2 \leftarrow$	$\mathbf{s_1}$ $\mathbf{s_2}?$	$\rightarrow r_2$ $\leftarrow r_2$	
t_3		$r_1 \rightarrow$ $r_2 \leftarrow$	$\mathbf{s_2}$ $\mathbf{s_2}?$	$\rightarrow r_2$ $\leftarrow r_2$	
t_4		$r_1 \rightarrow$ $r_2 \leftarrow$	$\mathbf{s_1}$ $\mathbf{s_2}?$	$\rightarrow r_2$ $\leftarrow r_2$	
t_5		$r_1 \rightarrow$ $r_2 \leftarrow$	$\mathbf{s_1}$ $\mathbf{s_2}?$	$\rightarrow r_2$ $\leftarrow r_2$	
t_6		$r_1 \rightarrow$ $r_2 \leftarrow$	$\mathbf{s_1}$ $\mathbf{s_1}?$	$\rightarrow r_2$ $\leftarrow r_2$	

Simulation

Method

We simulated 500 agent pairs for both models, for which code, data, and analyses can be found at: `https://osf.io/caeb9/`. The lexicon size was 4 (signals) $\times$ 3 (referents). Asymmetrical lexical bias was generated using a binomial distribution with $X = .45$ for agent 1 and $X = .55$ for agent 2. We set the rationality parameter α to 5 and, since irrelevant to our research focus, cost was set to 0. Each *dialogue* consisted of an initiating agent trying to communicate an intended referent to a responder agent. A dialogue took at most 6 *turns*, after which the pair gives up (given the limited lexicon, allowing for many or infinite turns will not provide further insight into the development of perceived and factual understanding). Furthermore, the agents performed six dialogues, with randomly determined intended referents.

Analysis 1: Clarification sequence length

We investigated the number of back-and-forth turns it takes for ostentively and non-ostentively communicating agents to come to the belief that they understand each other (called *clarification sequence length*). Fig. 4 shows a decrease in clarification sequence length after the first dialogue for both the ostensive and non-ostensive agents. This effect is consistent with other simulations and empirical observations that show an important role for interactive repair early in referential communication tasks (Hawkins et al., 2017; Degen, Trotzke, Scontras, Wittenberg, & Goodman, 2019; Mills, 2014). To get a better view of agents' factual understanding and of the distribution of clarification sequence lengths within dialogues, we perform more fine-grained analyses next.

Analysis 2: Dialogue

For each of the six dialogues that agent pairs engage in, we count the number of agent pairs that reach factual understanding (or not) and the number of turns they take, providing a distribution of repair sequence lengths (Fig. 5). Note that some agent pairs do not reach a state of perceived understanding and so give up after six turns (shaded areas in Fig. 5). While both models show a similar reduction in repair sequence length, this analysis reveals important differences in states of understanding for the two models.

First, ostensive agents achieve higher factual understanding overall. Here, factual understanding is formalized as the responder's inferred referent r being the same as the initiators intention i. Second, most ostensive agents reach a state of perceived understanding in the first half of the conversation (dialogues 1–3) and with short repair sequences (< 3). In contrast, many of the non-ostensive agents do not reach a state of perceived understanding with short repair sequences until the second half of the conversation (dialogues 4–6). Third, there is a substantial population of ostensive agents that in the final dialogue do not reach a state of perceived understanding and hence 'give up'. This seems to correspond accurately to the ratio of factual (mis)understandings they would have had, if they had to guess at this point. The non-ostensive agents show a different pattern: by the fifth or sixth intention almost all reach a state of perceived understanding but when they do, their average factual understanding is poor.

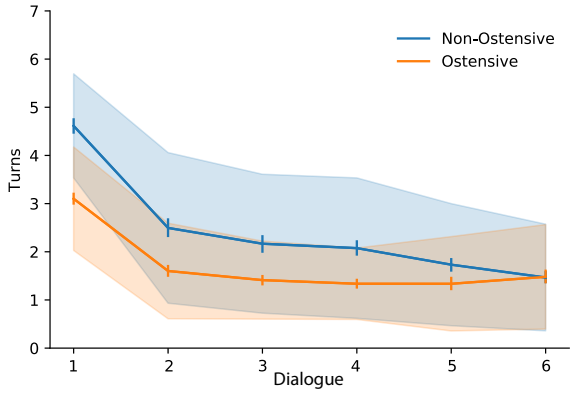


Figure 4: Clarification sequence length decreases as the conversation progresses for both ostensive (orange) and non-ostensive agents (blue). (Bars show the 95% CI and ribbon the SD).

This analysis shows that non-ostensive agents have a much harder time achieving factual understanding than ostensive agents. The ostensive agents can rely on the signal spoken by the initiator s_{init} and the ostensive response (e.g. pointing gesture) r by the responder. This provides them with a stable ground to infer the same referential meanings, which is only possible when their lexicon likelihoods align on a few compatible lexicons. The non-ostensive agents only have access to the signals s_{init} and s_{req} without the benefit of identification by pointing. They use pragmatic inference over all possible lexicons to infer not only the original intention underlying s_{init} but also the inferred intention of the responder's clarification request s_{req} (Eq. 7 and 8). While this additional inference leads to high certainty in a reasonable number of turns, it does not lead to factual understanding. This must mean that the non-ostensive agents fail to infer compatible lexicon likelihoods. In the next analysis we investigate the relationship between factual and perceived understanding.

Analysis 3: Factual and perceived understanding

We determined how many dialogues ended in factual understanding (see Table 2). Note that in the current simulation only 3 referents are available, making the chance level for

Table 2: Percentage of dialogues that end in factual understanding split by whether the agents perceived understanding or gave up. Chance level for factual understanding is 33.3%.

	perceived understanding	give up
ostensive	79.5 %	42.0 %
non-ostensive	34.4 %	33.4 %

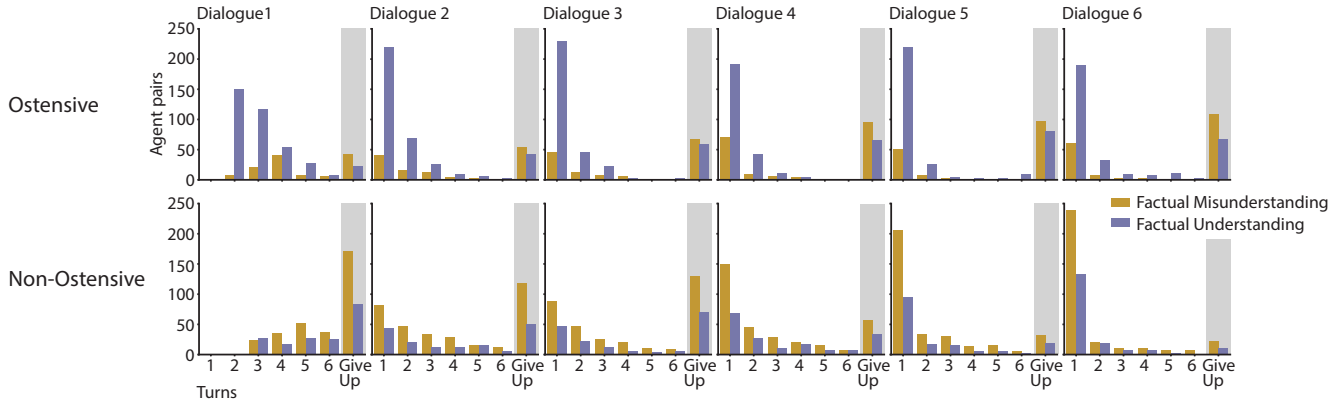


Figure 5: Structure of dialogue progression for ostensive (above) and non-ostensive agents (below). Conversation progresses from left to right. Each graph shows for the i^{th} dialogue the distribution of agent pairs over the repair sequence length. The rightmost bars in each graph shows how many agent pairs gave up in that dialogue and the colors indicate factual understanding.

factual understanding 33.3%. We observe again that when the ostensive agents reach a state of perceived understanding then they are above chance for achieving factual understanding, and when they give up they would have been below chance. The non-ostensive agents' factual understanding is always at chance level, suggesting that there is disconnect between their perceived and factual states of understanding.

Discussion

We formalized two models of communication: An ostensive model in which communicators can directly point to inferred referents to ask for clarification, and a non-ostensive model in which communicators can only rely on signals without direct access to referential identity. The agents in both models take conversation history into account—though each of a different nature—to try and infer a common lexicon to overcome differences in referential intention ascription between them.

Using agent-based simulations we showed that both models behave similar in one sense, namely, all agents become increasingly confident in their inferences leading to reduction in the number of turns taken per communicated intention. However, the models behave differently in other important ways. While ostensive agents have a high factual understanding when they perceive understanding, this is not the case for the non-ostensive agents. The non-ostensive agents' factual understanding is at chance level while at the same time also requiring more observations to gain enough certainty to stop asking for clarification. Though some degree of misalignment between factual and perceived understanding is fairly common in natural languages (Garfinkel, 1967; Enfield, 2007), these agents have no alignment at all.

The poor performance of the non-ostensive agents may seem surprising. While these agents lack physical co-presence (Clark & Marshall, 1981), they can rely on sophisticated inference and linguistic co-presence to reverse engineer possible meanings of (clarification) signals by considering all possible signal-referent mappings and their conversation history. This computationally far exceeds the inferential capacities of human communicators (van Rooij et al., 2011). Even so, these sophisticated inferences are not sufficient for the agents to achieve mutual (factual) understanding.

We started the paper by posing the challenge of explaining how people can make sense of a word when unable to point at a displaced referent. State-of-the-art models fail to address this challenge head-on by providing what we might call a Gavagai Shortcut: immediate access to the intended referent (Steels, 2015; Hawkins et al., 2017). This greatly simplifies the problem of converging on a common lexicon. However, the shortcut obscures the fact that this is interactionally and computationally far from trivial. The simulation results confirm that ostensive agents can achieve factual understanding, and they underscore the difficulty of computationally explaining non-ostensive communication. Without referentially clear signals, non-ostensive agents have no way of knowing when their inferences are factually correct. The challenge is clear: Without direct feedback, what computational infrastructure allows communicators to attain sufficient meta-understanding about their state of factual understanding? Given that the model presented here is not lacking in inferential capacity, it seems that more reasoning of the same kind is not the right answer. One might be tempted to introduce other shortcuts to 'make the current model work', but shortcuts forgo explanation. In future, computational theories of communication need to be expanded with a different kind of reasoning, one that explains how people can use context, background knowledge and other semiotic resources (Clark, 1997; Clark, 2006) to attain sufficient meta-understanding. Without this awareness, computational agents are not just trying to pull themselves from a swamp of underdetermination by their hair, but they don't realise they are drowning.

Acknowledgements

We would like to thank the anonymous reviewers for their invaluable feedback. LvdB is supported by a DCC PhD grant

awarded to MB, MD, IT and IvR. MB is supported by Netherlands Organization for Scientific Research (NWO) (Gravitation Grant 024.001.006 of the Language in Interaction consortium, LiI). MD is supported by NWO (016.Vidi.185.205). IvR acknowledges the support of a Distinguished Lorentz Fellowship funded by the Netherlands Institute for Advanced Study (NIAS) in the Humanities and Social Sciences and the Lorentz Center. IT is supported by NWO (453-08-002 and Gravitation Grant 024.001.006 of the LiI).

References

Bergen, L., & Goodman, N. D. (2015). The Strategic Use of Noise in Pragmatic Reasoning. *Topics in Cognitive Science*, *7*(2), 336–350.

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*.

Clark, A. (2006). Material Symbols. *Philosophical Psychology*, *19*(3), 291–307.

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

Clark, H. H. (1997). Dogmas of Understanding. *Discourse Processes*, *23*(3), 567–82.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In *Elements of Discourse Understanding.* Cambridge: Cambridge University Press.

Clark, H. H., & Schaefer, E. (1987, January). Collaborating on contributions to conversations. *Language and Cognitive Processes*, *2*(1), 19–41.

Cohn-Gordon, R., Goodman, N. D., & Potts, C. (2018). An incremental iterated response model of pragmatics. In *Proceedings of the Society for Computation in Linguistics (SCiL).* Linguistic Society of America.

Degen, J., Trotzke, A., Scontras, G., Wittenberg, E., & Goodman, N. D. (2019). Definitely, maybe: A new experimental paradigm for investigating the pragmatics of evidential devices across languages. *Journal of Pragmatics*, *140*, 33–48.

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., ... Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PLOS ONE*, *10*(9), e0136100.

Enfield, N. J. (2007). Tolerable friends. *Berkeley Linguistics Society*.

Frank, M. C., Emilsson, A. G., Peloquin, B., Goodman, N. D., & Potts, C. (2017). Rational speech act models of pragmatic reasoning in reference games. *PsyArXiv*. doi: 10.17605/OSF.IO/F9Y6B

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Garfinkel, H. (1967). Studies of the routine grounds of everyday activities. In H. Garfinkel (Ed.), *Studies in Ethnomethodology.* Englewood Cliffs, NJ: Prentice-Hall.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Grice, H. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics* (Vol. 3, pp. 41–58). New York, NY: Academic Press.

Hawkins, R. X. D., Frank, M. C., & Goodman, N. D. (2017). Convention-formation in iterated reference games. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Hawkins, R. X. D., Frank, M. C., & Goodman, N. D. (2020, April). Characterizing the dynamics of learning in repeated reference games. *arXiv:1912.07199 [cs]*.

Khani, F., Goodman, N. D., & Liang, P. (2018). Planning, inference and pragmatics in sequential language Games. *Transactions of the Association for Computational Linguistics*, *6*, 543–555.

Koivisto, A. (2015, February). Displaying Now-Understanding: The Finnish Change-of-State Token *aa*. *Discourse Processes*, *52*(2), 111–148. doi: 10.1080/0163853X.2014.914357

Mills, G. J. (2014). Dialogue in joint activity: Complementarity, convergence and conventionalization. *New Ideas in Psychology*, *32*, 158–173.

Quine, W. V. (1960). *Word and Object*. Cambridge: Cambridge University Press.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361–382.

Selting, M. (1987). Reparaturen und lokale Verstehensprobleme oder: Zur Binnenstruktur von Reparatursequenzen. *Linguistische Berichte*, *108*, 128–149.

Steels, L. (2015). *The Talking Heads experiment: Origins of words and meanings*. Language Science Press.

van Rooij, I., Kwisthout, J., Blokpoel, M., Szymanik, J., Wareham, T., & Toni, I. (2011). Intentional communication: Computationally easy or difficult? *Frontiers in Human Neuroscience*, *5*(52), 1–18.

Wilson, D., & Sperber, D. (2002). Relevance theory. In G. Ward & L. Horn (Eds.), *Handbook of pragmatics.* Blackwell.

Yoon, E. J., Frank, M. C., Tessler, M. H., & Goodman, N. D. (2018, December). *Polite speech emerges from competing social goals* (Tech. Rep.). PsyArXiv. doi: 10.31234/osf.io/67ne8