

Communicating with Speakers and Listeners of Different Pragmatic Levels

Kata Naszádi¹ and Frans A. Oliehoek² and Christof Monz¹

¹Language Technology Lab, University of Amsterdam

²Delft University of Technology

k.naszadi@uva.nl

Abstract

This paper explores the impact of variable pragmatic competence on communicative success through simulating language learning and conversing between speakers and listeners with different levels of reasoning abilities. Through studying this interaction, we hypothesize that matching levels of reasoning between communication partners would create a more beneficial environment for communicative success and language learning. Our research findings indicate that learning from more explicit, literal language is advantageous, irrespective of the learner's level of pragmatic competence. Furthermore, we find that integrating pragmatic reasoning during language learning, not just during evaluation, significantly enhances overall communication performance. This paper provides key insights into the importance of aligning reasoning levels and incorporating pragmatic reasoning in optimizing communicative interactions.

1 Introduction

In everyday conversations there is a trade-off between clarity and conciseness. Efficient messages might appear under-specified or ambiguous under a literal interpretation but can be successfully resolved using pragmatic reasoning about the speaker's intentions and the context of the communication (Grice, 1975; Horn, 1984; Fox and Katzir, 2011; Davies et al., 2022). If the speaker trusts the listener to make the right inferences, they can choose to be more concise. Being able to infer the intended meaning of an utterance beyond its literal content allows us to communicate efficiently.

The process of how people attain pragmatic interpretations using a model of the speaker's intentions has long been studied. There is also plenty of evidence from psycho-linguistic studies that individuals have different levels of pragmatic competence (Franke and Degen, 2016; Mayn et al., 2023). More importantly, people have been shown to keep track

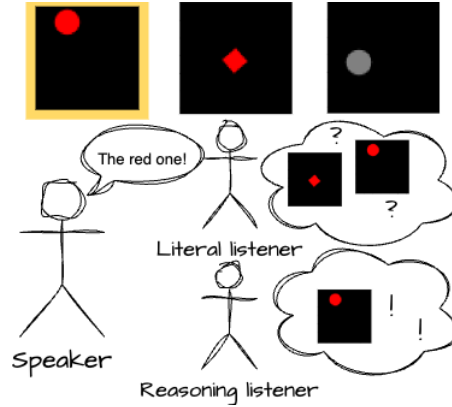


Figure 1: The speaker is asking for the red object. For a literal listener, this is ambiguous. A reasoning listener considers alternative messages about shape and color features and concludes that the speaker is asking for the red circle, as "square" would have been a more informative message for the other red object.

of the communicative partner's pragmatic competence and adjust their interpretations and messaging accordingly. This has been demonstrated both with human (Horton and Gerrig, 2002; Mayn et al., 2024) and artificial partners (Loy and Demberg, 2023; Branigan et al., 2011).

The pragmatic reasoning modeled in this work involves counterfactual reasoning about alternative sentences that the speaker could have uttered. The interaction in Figure 1 depicts an instance of such pragmatic reasoning about alternatives within our simple environment. According to pragmatic theory (Grice, 1975) the same process accounts for the interpretation "They are in the office for the rest of the week", when we hear the sentence "We are not in the office on Mondays".

In this work, we investigate the impact of varying pragmatic competence on communicative success. We pair literal and pragmatic listeners with speakers of different levels of pragmatic competence. We study the interaction between such speakers and listeners not only during inference, where both

partners have an already learned lexicon, but also during language learning. This way we gain insight into optimal levels of pragmatic inference for teachers and language learners. We hypothesise that matching levels of reasoning between partners benefits communicative success and language learning.

Our simulations reveal that with a lexicon that doesn't perfectly match that of the speaker's, sophisticated pragmatic listeners still significantly benefit from explicit literal language use. We also show that language learners that do not model pragmatic inference, struggle when learning from a speaker who uses pragmatic communication, while language learners that integrate a model of the speaker are significantly more successful.

2 Background

We situate our listener in an image-based version of Lewis's signaling game (Lewis, 1969). Image-referential games are commonly used to study the benefit of speakers and listeners reasoning about each other in context (Lee et al., 2018; White et al., 2020; Andreas and Klein, 2016).

At each turn a collection of N images is provided as context $C = (o_1, \dots, o_N)$, with the speaker having knowledge of a specific target image o_t , where $1 \leq t \leq N$. The listener's objective is to correctly identify the target image index t given the speaker's message w . The messages may contain multiple words by combining words from a fixed vocabulary.

2.1 Literal meanings and the Rational Speech Act model

Frank and Goodman (2012) provide a concise model for how speakers and listeners reason about each-other when sharing referential content. As a starting point, the model assumes an underlying literal interpretation. This is a function $D(w, o)$ of an utterance w and an observation o , in our case an image. In the original formulation the base interpretation function is a 0-1 valued indicator of the set of messages that are true of the image o . In line with other work, we replace this binary function with a real-valued similarity between the observed image-embedding and text-embedding.

$$D(o_i, w) = \text{CNN}_\theta(o_i)^T \text{RNN}_\theta(w) \quad (1)$$

Each image o_i is individually embedded with a CNN following the ResNet architecture (He et al.,

2016). The embedding of the message w is computed by an RNN with Gated Recurrent Units (Cho et al., 2014).

The listener models the distribution over the indices in an ordered set of images. The simplest listener distribution is produced by normalizing the score assigned by literal interpretation function over all the images in a given context C .

$$L_0(i|w, C) = \frac{e^{D(o_i, w)}}{\sum_{j=1}^{|C|} e^{D(o_j, w)}} \quad (2)$$

The speaker produces a message that maximizes the probability that the listener chooses the right image and also considers the cost of each message w . This means that the speaker has an internal model of the listener.

$$S_n(w|C, i) = \frac{e^{\lambda(\log(L_{n-1}(i|C, w)) - \text{cost}(w))}}{\sum_{w' \in V} e^{\lambda(\log(L_{n-1}(i|C, w')) - \text{cost}(w'))}} \quad (3)$$

In this work, we use a cost function that assigns a constant weight to each word and we only consider fully rational speakers with $\lambda = 1$. In the case of the speaker, the normalization happens over all possible messages $w \in V$. This is the most expensive step in the hierarchical reasoning process. In many natural language applications it is even prohibited by the fact that the set of all possible utterances is infinite. While exact inference is intractable, there are many papers discussing approximations (Cohn-Gordon et al., 2018; Liu et al., 2023; Lazaridou et al., 2020; White et al., 2020). In our communication-game, messages may contain one or two words: naming either the shape or the color of the target or both.

Building on 3, higher level listeners have an internal model of a speaker:

$$L_n(i|C, w) \propto S_{n-1}(w|C, i)P(C, i) \quad (4)$$

By applying Equations 3 and 4 in an alternating fashion, we can produce higher level speakers and listeners.

The most studied levels in the case of human communication are L_0 literal and L_2 pragmatic listeners paired with S_1 and S_3 speakers. This is motivated by evidence that humans can interpret messages from a S_3 speaker consistent with a L_2 listener (Goodman and Frank, 2016) and multiple pragmatic phenomena have been derived using the RSA framing and these levels (Franke and Degen, 2016; Hawkins et al., 2023).

2.2 Reasoning while learning

In the previous subsection 2.1 we saw how to perform recursive reasoning on top of given literal representations $D(o, w)$. These literal interpretations are most commonly initialized by functions learned outside of the context of a referential game and the reasoning is added only during inference (Fried et al., 2018; Lazaridou et al., 2020; Andreas and Klein, 2016; Liu et al., 2023).

However, the optimal literal representations are likely influenced by the reasoning itself. Following the work of Monroe and Potts (2015) and McDowell and Goodman (2019), we would like to integrate the knowledge that the received messages are the result of pragmatic reasoning already during learning. Therefore, we apply recursive reasoning during model training.

Pragmatic listeners seek to update the weights of the literal interpretation $D(o, w)$ but they need to do so by considering the repeated application of Equations 3 and 4. Similarly to McDowell and Goodman (2019), we derive the gradients of the reasoning process with respect to the lexicon weights. By repeated application of the chain rule through the hierarchical reasoning, pragmatic listeners backpropagate through the hierarchical reasoning and update the weights of the image- and utterance-embedding models.

3 Data

To investigate the impact of the pragmatic competence of speakers and listeners on communicative success, it is necessary to establish a controlled setting that allows for manipulation of the reasoning abilities of participants. We create a new environment based on the ShapeWorld dataset (Kuhnle and Copestake, 2017). Instead of the rule based method of Kuhnle and Copestake (2017), we use an exact implementation of the rational speaker defined in Equation 3. This way we can create speakers with different depth of recursive reasoning. Our speakers are not learned, they are knowledgeable users of the language: they have access to the underlying true lexicon which indicates the mapping between color and shape words and image properties.

Each game consists of a target image and a variable number of $N - 1$ distractor images. Images are described by one out of six different colors and a shape that can take five different values. The location, size and rotation of the objects is randomized on a 64x64 grid which creates a large variation of

candidate pictures.

We parameterize the process that generates the image tuples for each game by four probability distributions: the priors over the shapes $P(S)$ and colors $P(C)$, the probability that controls the correlations between colors $P(C|C)$ and the conditional defining the co-occurrence of shapes $P(S|S)$. We sample these distributions from different Dirichlet-distributions. We create two sets of concentration parameters: in the first version of the game, all sampled distributions are close to uniform ($Corr = 0$), while in the second version introduces correlations in the shape and color conditionals ($Corr = 1$). This way the sampled image tuples share more features, creating higher likelihood for pragmatic messaging that differentiates S_1 and S_3 .

For training, we sample only one instance of each distribution. At test time, we sample different $P(S)$, $P(C)$, $P(S|S)$ and $P(C|C)$ instances 10 times. From each of these constellations we sample 3200 games.

The random seed is fixed across all experiments and is reset for the learning and evaluation of each learner. This ensures that each listener sees the exact same examples in all environments.

4 Experiments

The fact that we have full control over the speaker’s messaging strategy and the data generating process allows us to alter the level of the speakers that the listeners learn from and create image tuples that highlight the contrast between higher level pragmatic and lower level literal messaging strategies.

We train L_0 literal listeners and L_2 pragmatic listeners. We create two different levels of speakers to pair them with our learning listeners: S_1 has an internal model of a competent L_0 , while S_3 anticipates L_2 -behavior.

Implementation for training and evaluating all models can be found at https://github.com/naszka/rsa_backward/.

4.1 Results

In this section, we present the insights gained from simulating language learning and communication between listeners and speakers with pragmatic or literal preferences. First we look at altering speaker and listener levels only during evaluation using an already trained lexicon. Then we turn to the learning dynamics between our four pairs: $L_0 - S_1$, $L_0 - S_3$, $L_2 - S_1$ and $L_2 - S_3$.

| Distractors | S_1 | S_3 |
|-------------|-------|-------|
| 2 | 1.07 | 1.01 |
| 3 | 1.14 | 1.02 |
| 4 | 1.24 | 1.09 |

Table 1: Average message length in words over 5000 samples for different number of distractors and speaker levels, $Corr = 1$. Higher level speakers send shorter messages and more distractors result in longer messages.

| | Listener eval | Speaker eval | Accuracy |
|----|---------------|--------------|----------|
| a) | 0 | 3 | 80.5 |
| b) | 2 | 3 | 81.2 ** |
| c) | 0 | 1 | 85.5 |
| d) | 2 | 1 | 85.6 |

Table 2: A listener trained as L_0 upgraded to different listener levels and paired with S_1 or S_3 at evaluation. Both L_0 and L_2 perform significantly better with the more verbose S_1 . When receiving messages from an S_3 , the higher level L_2 is significantly better. Evaluation setup: $cost = 0.6$, $N = 5$, $Corr = 1$.

Listening to speakers with different depth First we take the L_0 listener which learned in the easiest environment (S_1 , $Corr = 0$, $N = 3$) hence has the highest in-domain performance of 91.2% accuracy. During evaluation, we upgrade this listener to different levels: this means that during inference we apply recursive reasoning on top of the already learned L_0 lexicon. We pair these listeners with S_1 and S_3 . Table 2 shows that pragmatic L_2 is significantly ¹ better than literal L_0 when paired with S_3 . At the same time, L_2 still achieves the best performance with the more verbose S_1 , this is due to the fact that the listener did not learn the word-feature mapping with perfect accuracy and they still benefit from the more descriptive input.

We picked the evaluation parameters shown in Table 2 to maximize the speaker-type effect. The same trends hold for different number of distractors.

Learning from speakers with different depth Now we turn to how listeners of different levels are impacted by learning from different speakers.

Table 3 shows that reasoning learners that learned from lower level speakers always achieve higher accuracy at evaluation. This can be ex-

¹We perform Fisher’s exact test for significance testing. We note $p < 0.05$ with one asterisk * and for $p < 0.01$ we put ** next to the results.

| | Listener | Speaker train | Accuracy |
|----|----------|---------------|----------|
| a) | 0 | 1 | 80.7** |
| b) | | 3 | 79.1 |
| c) | 2 | 1 | 84.8** |
| d) | | 3 | 83.2 |

Table 3: For each level of listener, learning from lower level S_1 results in significantly better accuracy. Listener levels are kept the same during evaluation and training. Training and evaluation setup: $cost = 0.6$, $N = 5$, $Corr = 1$. Evaluation: S_1 .

plained by the fact that lower level speakers send longer messages on average, see Table 1, because their internal model is of a simpler listener who needs longer descriptions for success.

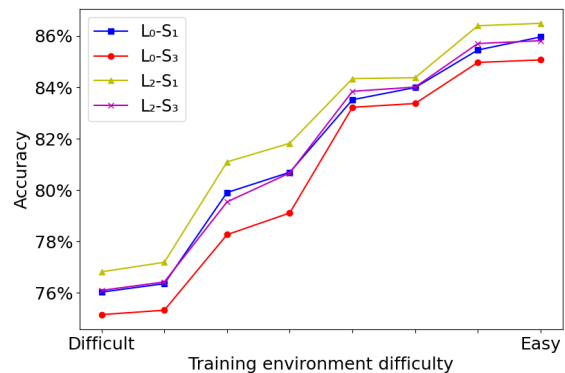


Figure 2: During training, listeners are paired with speakers of different pragmatic competence. The listeners are trained in environments of increasing difficulty. L_0 learners paired with S_1 speakers have the same performance as L_2 paired with S_3 .

Despite the fact that a L_2 can disambiguate S_3 messages, learning from a S_1 speaker is easier as it provides more data on both image features. This behaviour nicely aligns with the intuition that language learners benefit from simple, verbose communication and teachers should not assume challenging patterns of communicative competence early on in the learning process (Nguyen, 2022).

Comparing all possible pairings in Figure 2 however, we can clearly see the benefit of listeners having the appropriate level for the speaker during learning. A L_0 listener learning from a S_1 matches the performance of a L_2 listener learning from a S_3 speaker. We evaluate listeners that were paired with higher or lower level speakers during training. The evaluation environment is kept the same, all listeners are upgraded to L_2 and deployed with S_1 . Pragmatic L_2 listener can compensate for the dif-

faculty of learning from the concise S_3 through all training environments.

5 Conclusions

Humans exploit pragmatic reasoning in order to reduce the effort of speaking. For artificial agents to understand humans, it is critical to correctly resolve ambiguities. By recursively modeling the conversational partner, pragmatic listeners can arrive at the interpretations intended by pragmatic speakers.

In this work, we introduced speaker-listener pairs with matching or misaligned levels of pragmatic competence. We examined the benefits of integrating pragmatics not only during evaluation but already during language learning. Our results show that learning from more explicit, literal language is always beneficial, regardless of the pragmatic capacity of the learner. At the same time, we conclude that language learners need to apply reasoning about the context and the speaker when learning from data that was generated pragmatically.

6 Limitations

While the conversational phenomena we model in this paper have been widely attested to in linguistic theory and psycho-linguistic research, our experiments are limited to an artificial sandbox scenario with a small vocabulary and simple observations. The reasoning about all possible utterances used in this paper is intractable with larger vocabularies.

Real world conversations contain a wide range of pragmatic inferences, not all of which can be accounted for by the recursive reasoning presented in this paper.

7 Acknowledgements

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080. We also received funding from the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research with grant number 024.004.022.

References

Jacob Andreas and Dan Klein. 2016. [Reasoning about pragmatics with neural listeners and speakers](#). In *Proceedings of the 2016 Conference on Empirical Meth-*

ods in Natural Language Processing, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Holly P Branigan, Martin J Pickering, Jamie Pearson, Janet F McLean, and Ash Brown. 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, 121(1):41–57.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using RNN encoder–decoder for statistical machine translation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Reuben Cohn-Gordon, Noah D. Goodman, and Christopher Potts. 2018. [Pragmatically informative image captioning with character-level inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 439–443. Association for Computational Linguistics.

Catherine Davies, Vincent Porretta, Kremena Koleva, and Ekaterini Klepousniotou. 2022. Speaker-specific cues influence semantic disambiguation. *Journal of Psycholinguistic Research*, 51(5):933–955.

Danny Fox and Roni Katzir. 2011. On the characterization of alternatives. *Natural language semantics*, 19:87–107.

Michael C. Frank and Noah D. Goodman. 2012. [Predicting pragmatic reasoning in language games](#). *Science*, 336(6084):998–998.

Michael Franke and Judith Degen. 2016. Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one*, 11(5):e0154854.

Daniel Fried, Jacob Andreas, and Dan Klein. 2018. [Unified pragmatic models for generating and following instructions](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1951–1963, New Orleans, Louisiana. Association for Computational Linguistics.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Robert D Hawkins, Michael Franke, Michael C Frank, Adele E Goldberg, Kenny Smith, Thomas L Griffiths, and Noah D Goodman. 2023. From partners

- to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*, 130(4):977.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Laurence Horn. 1984. Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. In *Meaning, form, and use in context: Linguistic applications*, pages 11–42. Georgetown University Press.
- William S Horton and Richard J Gerrig. 2002. Speakers’ experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language*, 47(4):589–606.
- Alexander Kuhnle and Ann A. Copestake. 2017. Shape-world - a new test methodology for multimodal language understanding. *ArXiv*, abs/1704.04517.
- Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. [Multi-agent communication meets natural language: Synergies between functional and structural language learning](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online. Association for Computational Linguistics.
- Sang-Woo Lee, Yu-Jung Heo, and Byoung-Tak Zhang. 2018. Answerer in questioner’s mind: Information theoretic approach to goal-oriented visual dialog. *Advances in neural information processing systems*, 31.
- David Kellogg Lewis. 1969. *Convention: A Philosophical Study*. Cambridge, MA, USA: Wiley-Blackwell.
- Andy Liu, Hao Zhu, Emmy Liu, Yonatan Bisk, and Graham Neubig. 2023. [Computational language acquisition with theory of mind](#). In *The Eleventh International Conference on Learning Representations*.
- Ilya Loshchilov and Frank Hutter. 2017. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Jia E Loy and Vera Demberg. 2023. Perspective taking reflects beliefs about partner sophistication: Modern computer partners versus basic computer and human partners. *Cognitive Science*, 47(12):e13385.
- Alexandra Mayn, JE Loy, and Vera Demberg. 2023. Individual differences in overspecification: reasoning and verbal fluency.
- Alexandra Mayn, Jia E Loy, and Vera Demberg. 2024. Beliefs about the speaker’s reasoning ability influence pragmatic interpretation: Children and adults as speakers.
- Bill McDowell and Noah Goodman. 2019. [Learning from omission](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 619–628, Florence, Italy. Association for Computational Linguistics.
- Will Monroe and Christopher Potts. 2015. [Learning in the rational speech acts model](#). *CoRR*, abs/1510.06807.
- Minh Thi Thuy Nguyen. 2022. Interlanguage pragmatics as communicative competence. chapter 8, pages 135–151. Taylor & Francis.
- Julia White, Jesse Mu, and Noah D. Goodman. 2020. [Learning to refer informatively by amortizing pragmatic reasoning](#). In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 - August 1, 2020*. cognitivesciencesociety.org.

A Model training and implementation

All 261838 model-parameters are trained from scratch. The weights are updated with the AdamW optimizer (Loshchilov and Hutter, 2017) which we initialize with a learning rate of $1e - 5$.

For each training step, we use a batch of 32 games and the listeners are trained for 25920 training steps. Each instance of a listener training took 1.5 GPU hours on a single NVIDIA RTX A6000 GPU.

B Concentration parameters of the image generators

We sample $P(S)$, $P(C)$, $P(C|C)$ and $P(S|S)$ from Dirichlet distributions. In the case of no correlation between the images ($Corr = 0$), we set all concentration parameters to 1. For the correlated case ($Corr = 1$), we introduce correlation between the same shapes and a randomly chosen shape from all five shapes. We achieve this by setting the concentration parameter α to 5 at the index that corresponds to the i ’th shape and a randomly generated other index. $P(S|S = shape_i) \sim Dir(\alpha_1, \dots, \alpha_5)$, where all α ’s are 1 except for $\alpha_i = 5$ and $\alpha_j = 5$ for a randomly generated j . We apply the same process for generating all the $P(C|C)$ distributions.

C Benefits of pragmatic reasoning during learning

C.1 Pragmatic listeners learn faster

Figure 3 shows that when we keep all parameters of the learning environment constant, and only vary the listener’s depth, we observe that listeners with higher levels, learn to perform the task with good accuracy faster. The gap in performance is especially large in the initial learning stages. This result is in line with McDowell and Goodman (2019),

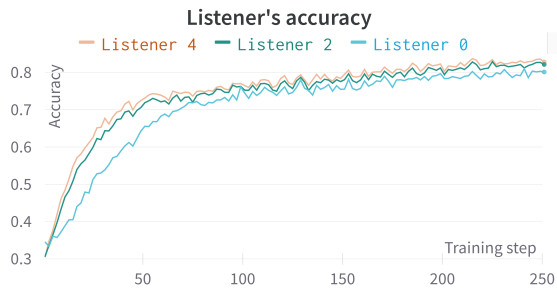


Figure 3: Higher level listeners learn quicker. In this comparison all other parameters such as speaker level, number of distractors, correlation between shapes are left constant.

where they discuss the benefits of pragmatic training.