# Shapeglot: Learning Language for Shape Differentiation

Panos Achlioptas* Judy Fan Robert Hawkins
Noah Goodman Leonidas Guibas
Stanford University

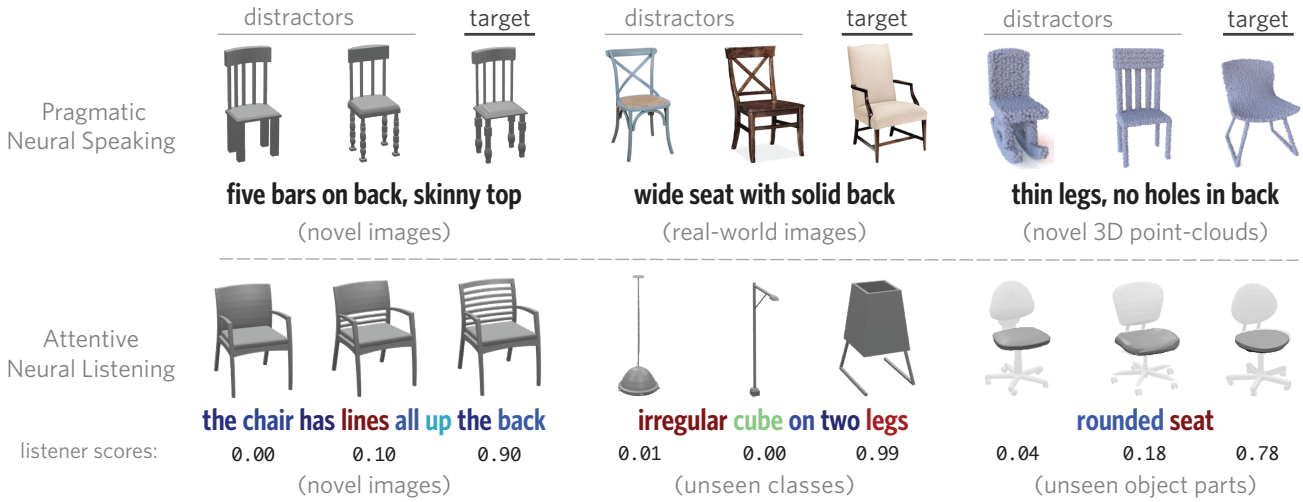



Figure 1: We introduce a novel corpus of utterances that refer to the shape of objects and use it to develop multimodal neural speakers and listeners with broad generalization capacity. **Top row**: Our neural speaker generates utterances to distinguish a 'target' shape from two 'distractor' shapes in unseen: images of synthetic data (left), *out-of-distribution* (OOD) real-world images (center), and 3D point-clouds of CAD models (right). **Bottom row**: Our neural listener interprets human-generated utterances in unseen (left-to-right): images of synthetic data, OOD object *classes* (here, lamps), and OOD isolated object *parts*. Listener scores indicate the model's confidence about which object the utterance refers to. The words are color-coded according to their importance, as judged by the attention module of this listener (warmer color indicates higher attention).

## Abstract

*In this work we explore how fine-grained differences between the shapes of common objects are expressed in language, grounded on 2D and/or 3D object representations. We first build a large scale, carefully controlled dataset of human utterances each of which refers to a 2D rendering of a 3D CAD model so as to distinguish it from a set of shape-wise similar alternatives. Using this dataset, we develop neural language understanding (listening) and production (speaking) models that vary in their grounding (pure 3D forms via point-clouds vs. rendered 2D images), the degree of pragmatic reasoning captured (e.g. speakers that reason about a listener or not), and the neural architecture (e.g. with or without attention). We find models that perform well with both synthetic and human partners, and with held out utterances and objects. We also find that these models are capable of zero-shot transfer learning to novel object classes (e.g. transfer from training on chairs to testing on lamps), as well as to real-world images drawn from furniture catalogs. Lesion studies indicate that the neural listeners depend heavily on part-related words and associate these words correctly with visual parts of objects (without any explicit supervision on such parts), and that transfer to novel classes is most successful when known part-related words are available. This work illustrates a practical approach to language grounding, and provides a novel case study in the relationship between object shape and linguistic structure when it comes to object differentiation.*

*Corresponding author: optas@cs.stanford.edu
Webpage: https://ai.stanford.edu/~optas/shapeglot

## 1. Introduction

Objects are best understood in terms of their structure and function, both of which rest on a foundation composed of object parts and their relations [9, 8, 47, 7]. Natural language has been optimized across human history to solve the problem of efficiently communicating the aspects of the world most relevant to one's current goals [19, 11]. As such, language can provide an effective medium to describe the *shape* and the *parts* of different objects, and as a result, to express *object differences*. For instance, when we see a chair we can analyze it into semantically meaningful parts, like its *back* and its *seat*, and can combine words to create utterances that reflect its geometric and topological *shape-properties* e.g. 'has a wide seat with a solid back'. Moreover, given a specific communication context, we can craft references that are not merely true, but which are also relevant e.g. we can refer to the lines found in a chair's back to *distinguish* it among other similar objects (see Fig. 1).

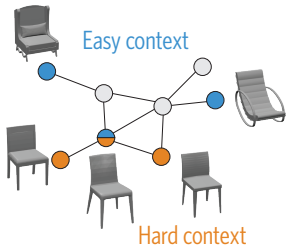
In this paper we explore this interplay between natural, referential language, and the shape of common objects. While a great deal of recent work has explored visually-grounded language understanding [18, 27, 44, 24, 23, 43], the resulting models have limited capacity to reflect the geometry and topology (i.e. the shape) of the underlying objects. This is because reference in previous studies was possible using properties like the object's *color*, or spatial configuration, including the absolute or relative (to other objects) *location*. Indeed, eliciting language that refers only to shape properties requires carefully controlling the objects, their presentation, and the underlying linguistic task. To address these challenges, in this work we utilize 3D CAD representations of objects which allow for flexible and *controlled* presentation (i.e. textureless, uniform-color objects, viewed in a fixed pose). We further make use of the 3D form to construct a reference game task in which the referred object is *shape-wise* similar to the distracting objects. The result of this effort is a new multimodal dataset, termed *ShapeGlot*, comprised of 4,511 unique chairs from ShapeNet [3] and 78,789 referential utterances. In ShapeGlot chairs are organized into 4,054 sets of size 3 (representing communication contexts) and each utterance is intended to distinguish a chair in context.

We use ShapeGlot to build and analyze a pool of modern neural language understanding (listening) and production (speaking) models. These models vary in their grounding (pure 3D forms via point-clouds vs. rendered 2D images), the degree of pragmatic reasoning captured (e.g. speakers that reason about a listener or not) and their precise neural architecture (e.g. with or without word attention, with context-*free*, or context-*aware* object encodings). We evaluate the effect of these choices on the original reference game task with both synthetic and human partners and find models with strong performance. Since language conveys abstractions such as object parts, that are shared between object categories, we hypothesize that our models learn robust representations that are *transferable* to objects of unseen classes (e.g. training on chairs while testing on lamps). Indeed, we show that these models have strong generalization capacity to novel object *classes*, as well as to *real-world* images drawn from furniture catalogs.

Finally, we explore *how* our models are succeeding on their communication tasks. We demonstrate that the neural listeners learn to prioritize the same abstractions in objects (i.e. properties of chair parts) that humans do in solving the communication task, despite *never* being provided with an explicit decomposition of these objects into parts. Similarly, we find that neural listeners transfer to novel object classes more successfully when known part-related words are available. Finally, we show that *pragmatic* neural speakers who consult an imagined (simulated) listener produce significantly more informative utterances than listener-unaware, *literal* speakers, as measured by human performance in identifying the correct object given the generated utterance.

## 2. Dataset and task

*ShapeGlot* consists of triplets of chairs coupled with referential utterances that aim to distinguish one chair (the 'target') from the remaining two (the 'distractors'). To obtain such utterances, we paired participants from Amazon's Mechanical Turk (AMT) to play an online reference game [15]. On each round of the game the two players were shown the same triplet of chairs. The designated target chair was privately highlighted for one player (the 'speaker') who was asked to send a message through a chat box such that their partner (the 'listener') could successfully select it. To ensure speakers used *only* shape-related information, we scrambled the positions of the chairs for each participant independently and used textureless, uniform-color renderings of pre-aligned 3D CAD models, taken from the same viewpoint. To ensure that the communicative interaction was natural, no constraints were placed on the chat box: referring expressions from the speaker were occasionally followed by clarification questions from the listener or other discourse.




A key decision in building our dataset concerned the construction of contexts that would reliably elicit *diverse* and potentially *very* fine-grained contrastive language. To achieve diversity we considered all ∼7,000 chairs from ShapeNet. This object class is geometrically complex, highly diverse, and abundant in the real world. To control

the granularity of fine-grained distinctions that were necessary in solving the communication task, we constructed two types of contexts: *hard* contexts consisted of very similar shape-wise chairs, and *easy* contexts consisted of less similar chairs. To measure shape-similarity in a scalable manner, we used the semantically rich latent space of a Point Cloud-AutoEncoder (PC-AE) [1]. We note, that point-clouds are an intrinsic representation of a 3D object, *oblique* to color or texture. After extracting a 3D point-cloud from the surface of each ShapeNet model we computed the underlying K-nearest-neighbor graph among all models according to their PC-AE embedding distances. For a chair with sufficiently high-in degree on this graph (corresponding intuitively to a canonical chair) we contrasted it with four distractors: the two *closest* to it in latent-space, and two that were sufficiently far (see inset and Supplementary Materials for details). Last, we note that we *counterbalanced* the collected utterances, by considering every chair in a given context as the context's target (in different games).

Before we present our neural agents, we identify some distinctive properties of our corpus. Human performance on the reference game was high, but listeners made significantly more errors in the hard contexts (accuracy $94.2\%$ vs. $97.2\%, z = 13.54, p < 0.001$). Also, in hard contexts longer utterances were used to describe the targets (on average 8.4 words vs. 6.1, $t = -35, p < 0.001$). A wide spectrum of descriptions was elicited, ranging from more holistic/categorical (e.g. 'the rocking chair') common for easy contexts, to more complex and fine-grained language, (e.g. 'thinner legs but without armrests') common for hard ones. Interestingly, 78% of the produced utterances contained at least one part-related word: *back*, *legs*, *seat*, *arms*, or closely related synonyms e.g. *armrests*.

## 3. Neural listeners

Developing neural listeners that reason about shape-related properties is a key contribution of our work. Below we conduct a detailed comparison between three distinct architectures, highlight the effect of different regularization techniques, and investigate the merits of different representations of 3D objects for the listening task, namely, 2D rendered images and 3D surface point clouds. In what follows, we denote the three objects of a communication context as $O = \{o_1, o_2, o_3\}$, the corresponding word-tokenized utterance as $U = u_1, u_2, \ldots$ and as $t \in O$ the designated target.

Our proposed listener is inspired by [26]. It takes as input a (latent code) vector that captures shape information for each of the objects in $O$, and a (latent code) vector for each token of $U$, and outputs an *object–utterance* compatibility score $\mathcal{L}(o_i, U) \in [0, 1]$ for each input object. At its core lies a multi-modal LSTM [16] that receives as initial input ("is grounded" with) the vector corresponding to one object, processes the word-sequence $U$, and is read out by an MLP to yield a single number (a compatibility score). This is *repeated* for each object, while *sharing* all network parameters across the objects. The resulting three scores are soft-max normalized and compared to the ground-truth indicator vector of the target, under the cross-entropy loss.*

**Shape encoders** We experiment with three representations to capture the shapes of the underlying objects: (a) the bottleneck vector of a pretrained Point Cloud-AutoEncoder (PC-AE), (b) the embedding provided by a convolutional network operating on single-view images of non-textured 3D objects, or (c) a combination of (a) and (b). Specifically, for (a) we use the PC-AE architecture of [1] trained with single-class point clouds extracted from the surfaces of 3D CAD models, while for (b) we use the activations of the penultimate layer of a VGG-16 [32], pre-trained on ImageNet [6], and fine-tuned on an 8-way classification task with images of objects from ShapeNet. For each representation we project the corresponding latent code vector to the input space of the LSTM using a fully connected (FC) layer with $L_2$-norm weight regularization. While there are many ways to incorporate image-based with point-cloud based features in the LSTM, we found that the best performance occurs when we i) ground the LSTM with the image-based codes, ii) concatenate the LSTM's output (after processing $U$) with the point cloud-based codes, and iii) feed the concatenated result in a shallow MLP that produces the compatibility score (see Supp. for a *visual overview* of the pipeline and more details). We note that proper regularization is *critical*: adding dropout at the input layer of the LSTM and $L_2$ weight regularization and dropout at and before the FC projecting layers improves performance $\sim 10\%$.

**Incorporating context information** Our baseline listener architecture (*Baseline*, just described) first scores each object *separately* then applies softmax normalization to yield a score distribution over the three objects. We also consider two alternative architectures that explicitly encode information about the *entire* context before scoring a single object. The first alternative (*Early-Context*), is identical to the proposed architecture, except for the codes used to *ground* the LSTM. Specifically, if $v_i$ is the image-based code vector of the i-th object, instead of using $v_i$ as the grounding vector for $o_i$, a shallow convolutional network is introduced to create a more complex (context-aware) feature. This network, of which the output *is* the grounding code for $o_i$, receives the signal $f(v_j, v_k)||g(v_j, v_k)||v_i$, where $f, g$ are the symmetric max/mean-pool functions, $||$ denotes feature-wise concatenation and $v_j, v_k$ are the codes of the remaining objects. Here, we use symmetric functions to induce the orderless nature of our contexts. The second alternative (*Combined-Interpretation*) inputs the image-based code vectors for *all* three objects se-

*Architecture details and hyper-parameters for all the experiments, are provided in the Supplementary Materials.

quentially to the LSTM and then proceeds to process the tokens of $U$ *once*, before yielding three scores. Similarly to the *Baseline* architecture, point clouds are incorporated in both alternatives at the MLP operating after the LSTM.

**Word attention** We hypothesized that a listener forced to prioritize a few tokens in each utterance would learn to prioritize tokens that express properties that distinguish the target from the distractors (and, thus, perform better). To test this hypothesis, we augment the listener models with a standard *bilinear attention mechanism* [31]. Specifically, to estimate the 'importance' of each token $u_i$ we compare the output of the LSTM when it inputs $u_i$ (denoting the output as $r_i$); with the hidden state *after* the entire utterance has been processed (denoted as $h$). The relative importance of each token is $a_i \triangleq r_i^T \times W_{\text{att}} \times h$, where $W_{\text{att}}$ is a trainable diagonal matrix. The new (weighted) output of the LSTM is: $\sum_{i=1}^{|U|} r_i \odot \hat{a}_i$, where $\hat{a}_i = \frac{\exp(a_i)}{\sum_j^{|U|} \exp(a_j)}$ and $\odot$ is the point-wise product.

## 4. Listener experiments

We begin our evaluation of the proposed listeners using two reference tasks based on different data splits. In the *language generalization* task, we test on target objects that were seen as targets in at least one context during training but ensure that all utterances in the test split are from unseen speakers. In the more challenging *object generalization* task, we restrict the set of objects that appeared as targets in the test set to be *disjoint* from those in training such that all speakers *and* objects in the test split are new. For each of these tasks, we evaluate choices of input modality and word attention, using $[80\%, 10\%, 10\%]$ of the data, for training, validating and testing purposes.

*Baseline* listener accuracies are shown in Table 2.[†] Overall the *Baseline* achieves good performance. As expected, the listeners have higher accuracy on the language generalization task ($3.2\%$ on average). The attention mechanism on words yields a mild performance boost, as long as images are part of the input. Interestingly, images provide a significantly better input than point-clouds when only one modality is used. This may be due to the higher-frequency content of images (we use point-clouds with only 2048 points), or the fact that VGG was pre-trained while the PC-AE was not. However, we find *significant* gains in accuracy ($4.1\%$ on average) from exploiting the two object representations *simultaneously*, implying a complementarity among them.

Next, we evaluate how the different approaches in incorporating context information described in Section 3 affect listener performance. We focus on the more challenging object generalization task, using listeners that include attention and both object modalities. We report the findings in Table 1. We find that the *Baseline* and *Early-Context* models perform best overall, outperforming the *Combined-Interpretation* model, which does not share weights across objects. This pattern held for both hard and easy contexts of our dataset. We further explore the small portion ($\sim 14\%$) of our test set that use explicitly contrastive language: superlatives ('skinniest') and comparatives ('skinnier'). Somewhat surprisingly we find that the *Baseline* architecture remains competitive against the architectures with more explicit context information. The *Baseline* model thus achieves high performance and is the most flexible (at test time it can be applied to *arbitrary-sized* contexts); we focus on this architecture in the explorations below.

### 4.1. Exploring learned representations

**Linguistic ablations** Which aspects of a sentence are most critical for our listener's performance? To inspect the properties of words receiving the most attention, we ran a part-of-speech tagger on our corpus. We found that the highest attention weight is placed on *nouns*, controlling for the length of the utterance. However, adjectives that *modify* nouns received more attention in hard contexts (controlling for the average occurrence in each context), where nouns are often not sufficient to disambiguate (see Fig. 2A). To more systematically evaluate the role of higher-attention tokens in listener performance, we conducted an utterance lesioning experiment. For each utterance in our dataset, we successively replaced words with the `<UNK>` token according to three schemes: (1) from highest attention to lowest, (2) from lowest attention to highest, and (3) in random order. We then fed these through an equivalent listener trained *without* attention. We found that up to 50% of words can be removed without much performance degradation, but only if these are low attention words (see Fig. 2B). Our word-attentive listener thus appears to rely on context-appropriate content words to successfully disambiguate the referent.

**Visual ablations** To test the extent to which our listener is relying on the same semantic *parts* of the object as humans, we next conducted a lesion experiment on the visual input. We took the subset of our test set where (1) all chairs had complete part annotations available [42] and (2) the corresponding utterance mentioned a *single* part (17% of our test set). We then created lesioned versions of all three objects on each trial by removing pixels of images (and/or points when point-clouds are used), corresponding to parts according to two schemes: *removing* a single part or *keeping* a single part. We did this either for the mentioned one, or another part, chosen at random. We report listener accuracies on these lesioned objects in Table 3. We found that removing random parts hurts the accuracy by 10.4% on average, but removing the mentioned part dropped accuracy more than

[†]In all results mean accuracies and standard errors across 5 random seeds are reported, to control for the data-split populations and the initialization of the neural-network.

| Architecture | Overall | Subpopulations | | |
|---|---|---|---|---|
| | | Hard | Easy | Sup-Comp |
| *Combined-Interpretation* | $75.9 \pm 0.5\%$ | $67.4 \pm 1.0\%$ | $83.8 \pm 0.6\%$ | $74.4 \pm 1.5\%$ |
| *Early-Context* | $79.4 \pm 0.8\%$ | $\mathbf{70.1} \pm 1.3\%$ | $88.1 \pm 0.6\%$ | $75.6 \pm 2.2\%$ |
| *Baseline* | $\mathbf{79.6} \pm 0.8\%$ | $69.9 \pm 1.3\%$ | $\mathbf{88.8} \pm 0.4\%$ | $\mathbf{76.3} \pm 1.3\%$ |

Table 1: Comparing different ways to include context. The simplest *Baseline* model performs as well as more complex alternatives. Subpopulations are the subsets of test data containing: hard contexts (shape-wise similar distractors), easy contexts, superlatives or comparatives.

| | Input Modality | Language Task | Object Task |
|---|---|---|---|
| **No Attention** | Point Cloud | $67.6 \pm 0.3\%$ | $66.4 \pm 0.7\%$ |
| | Image | $81.2 \pm 0.5\%$ | $77.4 \pm 0.7\%$ |
| | Both | $83.1 \pm 0.4\%$ | $78.9 \pm 1.0\%$ |
| **With Attention** | Point Cloud | $67.4 \pm 0.3\%$ | $65.6 \pm 1.4\%$ |
| | Image | $81.7 \pm 0.5\%$ | $77.6 \pm 0.8\%$ |
| | Both | $\mathbf{83.7} \pm 0.3\%$ | $\mathbf{79.6} \pm 0.8\%$ |

Table 2: Performance of the *Baseline* listener architecture using different object representations and with/without word level attention, in two reference tasks.

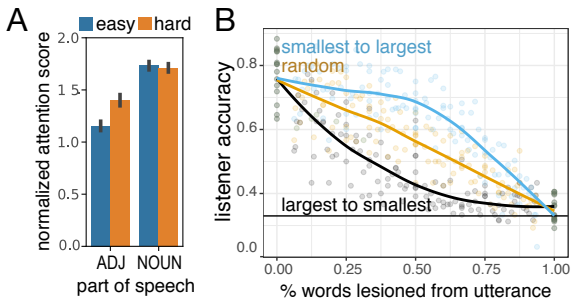



Figure 2: (A) The listener places more attention on adjectives in hard (orange) triplets than easy (blue) ones. The histogram's heights depict mean attention scores normalized by the length of the underlying utterances; the error bars are bootstapped 95% confidence intervals. (B) Lesioning highest attention words to lowest worsens performance more than lesioning random words or lesioning lowest attention words.

three times as much, nearly to chance. Conversely, keeping *only* the mentioned part while lesioning the rest of the image merely drops accuracy by 10.6% while keeping a non-mentioned (random) part alone brings accuracy down close to chance. In other words, on trials when participants depended on information about a part to communicate the object to their partner, we found that visual information about that part was both *necessary and sufficient* for the performance of our listener model.

| | Single Part Lesioned | Single Part Present |
|---|---|---|
| **Mentioned Part** | $42.8\% \pm 2.3$ | $66.8\% \pm 1.4$ |
| **Random Part** | $67.0\% \pm 2.9$ | $38.8\% \pm 2.0$ |

Table 3: Evaluating the part-awareness of neural listeners by lesioning object *parts*. Results shown are for image-only listeners, with average accuracy of $77.4\%$ when *intact* objects are used. Similar findings regarding point-cloud-based listeners are provided in the Supplementary Materials.

## 5. Neural speakers

**Architecture** Next, we explore models that learn to generate an utterance that refers to the target and which distinguishes it from the distractors. Similarly to a neural listener the heart of these (speaker) models is an LSTM which encodes the objects of a communication context, and then decodes an utterance. Specifically, for an *image-based* speaker, on the first three time steps, the LSTM input is the VGG code of each object. Correspondingly, for a *point-cloud-based* speaker, the LSTM's initial input is the object codes extracted from a PC-AE. During training and after the object codes are processes by the LSTM, the LSTM receives sequentially the $i$-th utterance token, while at its output if forced to predict the $(i+1)$-th token (i.e. we use teacher-force [38]). For these models we feed the target object always last (third), eliminating the need to represent an index indicating the target's position. To find the best model hyper-parameters (e.g. $L_2$-weights, dropout-rate and # of LSTM neurons) and the optimal amount of training, we sample synthetic utterances from the model during training and use a pretrained *listener* to select the result with the highest listener accuracy. We found this approach to produce results that yield better quality utterances than evaluating with listening-unaware metrics like BLEU [29].

**Variations** The above (*literal*) speakers can learn to generate language that discriminates targets from distractors. To test the degree to which distractor objects are used for this purpose, we experimented with *context-unaware* speakers that were provided with the latent code of the target *only*,

(and are otherwise identical to the above *literal* models). Furthermore, and motivated by the recursive social reasoning characteristic of human pragmatic language use (as formalized in the Rational Speech Act framework [12]), we created *pragmatic* speakers that choose utterances according to their capacity to be discriminative, as judged by a pretrained 'internal' listener. In this case, we sample utterances from the (*literal*) speakers, but score (i.e. re-rank) them with:

$$\beta \log(P_L(t|U,O)) + \frac{(1-\beta)}{|U|^\alpha} \log(P_S(U|O,t)), \qquad (1)$$

where $P_L$ is the listener's probability to predict the target ($t$) and $P_S$ is the likelihood of the *literal* speaker to generate $U$. The parameter $\alpha$ controls a length-penalty term to discourage short sentences [40], while $\beta$ controls the relative importance of the speaker's vs. the listener's opinions.

## 6. Speaker experiments

Qualitatively, our speakers produce good object descriptions, see Fig. 3 for examples, with the pragmatic speakers yielding more discriminating utterances.[‡] To quantitatively evaluate the speakers we measured their success in reference games with two different kinds of partners: with independently-trained neural listeners and with human listeners. To conduct a *fair* study when we used a neural listener for evaluations, we split the training data in half. The evaluating listener was trained using one half, while the 'internal' listener used by the pragmatic speaker was trained on the remaining half. For the human-based evaluations, we first used the *literal* and *pragmatic* variants to generate an utterance for every context of the test split of the *object-generalization* task (which contains 1200 unique contexts). We then showed the resulting utterances to participants recruited with AMT and asked them to select the object from context that the speaker was referring to. We collected approximately 2.2 responses for each context. Here, we used the synthetic utterances with the highest scores (Eq. 1) from each model, with optimal (per-validation) $\alpha$ and an 'aggressive' $\beta = 1.0$. We note that while the *point-based* speakers operate *solely* with 3D point-clouds, we sent their generated utterances to AMT coupled with CAD rendered images, so as to keep the visual (AMT-human) presentation identical across the two variants.

We found (see Table 4) that our *pragmatic* speakers perform best with both neural and human partners. While their success with the neural listener model may be unsurprising, given the architectural similarity of the internal listener and the evaluating listener, *human* listeners were 10.4 percentage points better at picking out the target on utterances produced by the *pragmatic* vs. *literal* speaker for the best-performing (*image-based*) variant. Similar to what we saw in the listener experiments (Section 4), we found that (sole) point-cloud-based speakers achieve lower performance than image-based variants. However, we also found an asymmetry between the listening and speaking tasks: while context-unaware (*Baseline*) listeners achieved high performance, we found that context-unaware speakers fare significantly worse than context-aware ones. Last, we note that both literal and pragmatic speakers produce *succinct* descriptions (average sentence length 4.21 vs. 4.97) but the pragmatic speakers use a much richer vocabulary (14% more unique nouns and 33% more unique adjectives, after controlling for average length discrepancy).

[‡]The project's webpage contains additional qualitative results.

Table 4: Evaluating neural speakers operating with 3D point-cloud or image-based object representations, across architectural variants.

| Speaker Architecture | Modality | Neural Listener | Human Listener |
|---|---|---|---|
| Context Unaware | Point Cloud | $59.1 \pm 2.0\%$ | - |
| | Image | $64.0 \pm 1.7\%$ | - |
| Literal | Point Cloud | $71.5 \pm 1.3\%$ | 66.2 |
| | Image | $76.6 \pm 1.0\%$ | 68.3 |
| Pragmatic | Point Cloud | $90.3 \pm 1.3\%$ | 69.4 |
| | Image | **92.2** $\pm 0.5\%$ | **78.7** |

## 7. Out-of-distribution transfer learning

Language is abstract and compositional. These properties make language use generalizable to new situations (e.g. using concrete language in novel scientific domains) and robust to low-level perceptual variation (e.g. lighting). In our final set of experiments we examine the degree to which our neural listeners and speakers learn representations that are correspondingly *robust*: that capture associations between the visual and the linguistic domains that permit generalization out of the training domain.

**Understanding out-of-class reference** To test the generalization of listeners to novel stimuli, we collected referring expressions in communication contexts made of objects in ShapeNet drawn from new classes: beds, lamps, sofas and tables. These classes are distinct from chairs, but share some parts and properties, making transfer possible for a sufficiently compositional model. For each of these classes we created 200 contexts made of random triplets of objects; and collected 2 referring expressions for each target in each context (from participants on AMT). Examples of visual stimuli and collected utterances are shown in Fig. 4 (bottom-row). To this data, we applied an (image-only, with/without-attention) listener trained on the ShapeGlot (i.e. chairs) data. We avoid using point-clouds since

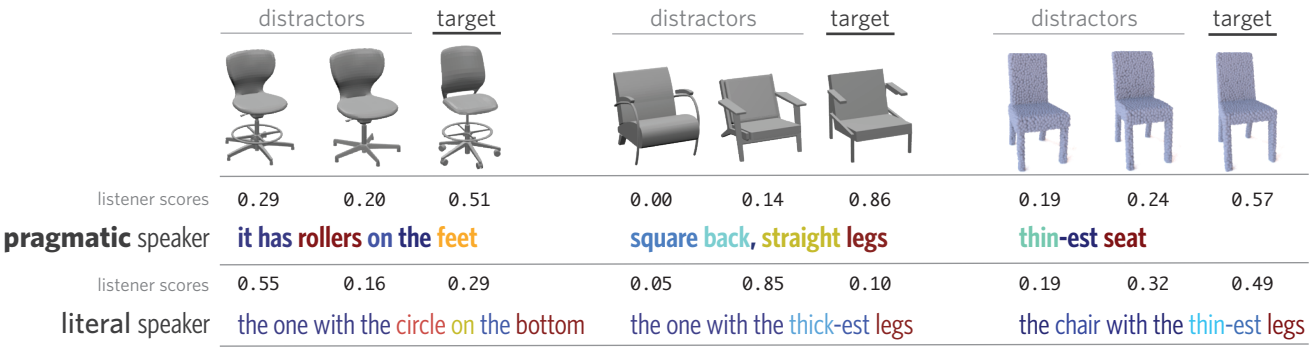



Figure 3: *Pragmatic* vs. *literal* speakers in unseen ('hard') contexts. The pragmatic generations successfully discern the target even in cases where the literal generations fail. The left and center contexts (gray-color) are used by image-based speakers/listeners, and the right-most by point-cloud-based ones. The utterances are color-coded according to the attention placed by a separate evaluating neural listener whose classification scores are shown above each corresponding utterance.

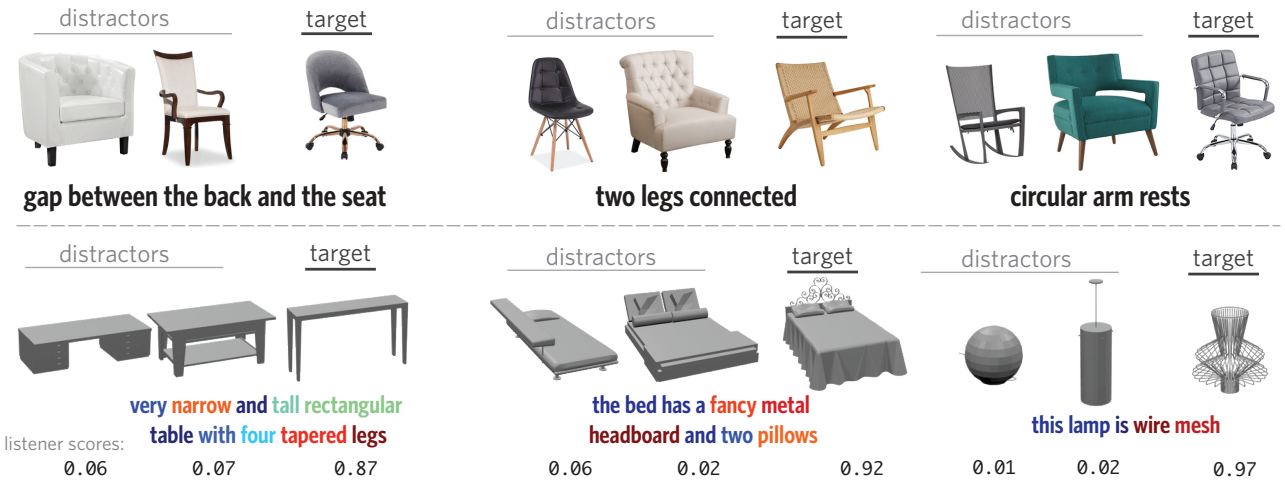



Figure 4: Examples of *out-of-distribution* neural speaking and listening. **Top row**: model generations for *real-world* catalogue images. The speaker successfully describes fine grained shape differences on images with rich color and texture content; two factors *not* present in the training data. **Bottom row**: results of applying a word-attentive listener on renderings of CAD objects from *unseen* classes with human-produced utterances. The listener can detect the (often localized) visual cues that humans refer to, despite the large visual discrepancy of these objects from the training-domain of chairs. (The utterances are color coded according to the attention placed to them by the attentive neural listener.)

unlike VGG which was finetuned with multiple ShapeNet classes, the PC-AE was pre-trained on a single-class.

As shown in Table 5, the average accuracy is well above chance in all transfer categories (56% on average). Moreover, constraining the evaluation to utterances that contain *only* words that are in the ShapeGlot training vocabulary (75% of all utterances, column: *known*) only slightly improves the results. This is likely because utterances with unknown words still contain enough known vocabulary for the model to determine meaning. We further dissect the *known* population into utterances that contain part-related words (*with-part*) and their complement (*without-part*). For the training domain of chairs without-part utterances yield slightly higher accuracy. However the useful subcategories that support this performance (e.g. 'recliner') do not support transfer to new categories. Indeed, we observe that for transfer classes the listener performs better when part-related words are present. Furthermore, the performance gap between the two populations appears to become larger as the perceptual distance between the transfer and training domains increases (compare sofas to lamps).

**Describing real images** Transfer from synthetic data to real data is often difficult for modern machine learning models, that are attuned to subtle statistics of the data. We explored the ability of our models to transfer to real chair images (rather than the training images which were ren-

Table 5: Transfer-learning of neural listeners trained with chair data to novel object classes for different subpopulations of utterances. For reference, the accuracies of the *object generalization* task are included (chairs, first row); The last row reports the average of the transfer/novel categories only. All numbers are *average* accuracies of five listeners trained with different splits of the *object generalization* task (See Section 7 for details, and Supp. for other variants.).

| | Population | | | |
|---|---|---|---|---|
| Class | entire | known | with part | without part |
| chair | 77.4 | 77.8 | 77.0 | **80.5** |
| bed | 56.4 | 55.8 | **63.8** | 51.5 |
| lamp | 50.1 | 51.9 | **60.3** | 47.1 |
| sofa | 53.6 | 55.0 | **55.1** | 54.7 |
| table | 63.7 | 65.5 | **68.3** | 62.7 |
| average | 56.0 | 57.1 | **61.9** | 54.9 |

dered without color or texture from CAD models) by curating a modest-sized (300) collection of chair images from online furniture catalogs. These images were taken from a *similar* view-point to that of the training renderings and have rich color and texture content. We applied the (image-only) *pragmatic* speaker to these images, after subtracting the average ImageNet RGB values (i.e. before passing the images to VGG). Examples of the speaker's productions are shown in Figure 4. For each chair, we randomly selected two distractors and asked 2 AMT participants to guess the target given the (highest-scoring) utterance produced by our speaker. Human listeners correctly guessed the target chair $70.1\%$ of the time. Our speaker appears to transfer successfully to real images, which contain color, texture, pose variation, and likely other differences from our training data.

## 8. Related work

**Image labeling and captioning** Our work builds on recent progress in the development of vision models that involve some amount of language data, including object categorization [32, 46] and image captioning [17, 37, 41]. Unlike object categorization, which pre-specifies a fixed set of class labels to which all images must project, our systems use open-ended, referential language. Similarly to other recent works in image captioning [25, 27, 44, 35, 24, 23, 43], instead of captioning a single image (or entity therein), in isolation, our systems learn how to communicate across diverse communication contexts.

**Reference games** In our work we use reference games [18] in order to operationalize the demand to be relevant in context. The basic arrangement of such games can be traced back to the language games explored by Wittgenstein [39] and Lewis [22]. For decades, such games have been a valuable tool in cognitive science to quantitatively measure inferences about language use and the behavioral consequences of those inferences [30, 20, 4, 34]. Recently, these approaches have also been adopted as a benchmark for discriminative or context-aware NLP [28, 2, 33, 36, 26, 5, 21].

**Rational speech acts framework** Our models draw on recent formalization of human language use in the Rational Speech Acts (RSA) framework [12]. At the core of RSA is the Gricean proposal [14] that speakers are agents who select utterances that are parsimonious yet informative about the state of the world. RSA formalizes this notion of informativity as the expected reduction in the uncertainty of an (internally simulated) listener, as our pragmatic speaker does. The literal listener in RSA uses semantics that measure compatibility between an utterance and a situation, as our baseline listener does. Previous work has shown that RSA models account for context sensitivity in speakers and listeners [13, 26, 45, 10]. Our results add evidence for the effectiveness of this approach in the shape domain.

## 9. Conclusion

In this paper, we have explored models of natural language grounded in the shape of common objects. The geometry and topology of objects can be complex and the language we have for referring to them is correspondingly abstract and compositional. This makes the shape of objects an ideal domain for exploring grounded language learning, while making language an especially intriguing source of evidence for shape variations. We introduced the *ShapeGlot* corpus of highly descriptive referring expressions for shapes in context. Using this data we investigated a variety of neural listener and speaker models, finding that the best variants exhibited strong performance. These models draw on both 2D and 3D object representations and appear to reflect human-like part decomposition, though they were never explicitly trained with object parts. Finally, we found that the learned models are surprisingly robust, transferring to real images and to new classes of objects. Future work will be required to understand the transfer abilities of these models and how this depends on the compositional structure they have learned.

**Acknowledgements:** The authors wish to thank all the anonymous reviewers for their insightful comments and suggestions. P.A. wishes to also thank Dimitris Achlioptas for numerous inspiring conversations. Last but not least, the authors would like to acknowledge support from the NSF grant CHS-1528025, a Vannevar Bush Faculty Fellowship, and gifts from Amazon Web Services for Machine Learning Research and Autodesk.

## References


[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas J. Guibas. Learning representations and generative models for 3d point clouds. *Proceedings of the 35th International Conference on Machine Learning*, 2018. 3

[2] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. *CoRR*, 2016. 8

[3] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015. 2

[4] Herbert H Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986. 8

[5] Reuben Cohn-Gordon, Noah Goodman, and Chris Potts. Pragmatically informative image captioning with character-level reference. *CoRR*, abs/1804.05417, 2018. 8

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[7] Anastasia Dubrovina, Fei Xia, Panos Achlioptas, Mira Shalah, and Guibas J. Leonidas. Composite shape modeling via latent space factorization. *CoRR*, abs/1901.02968, 2019. 2

[8] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010. 2

[9] A. Martin Fischler and Elschlager A. Robert. The representation and matching of pictorial structures. *IEEE Trans. on Computers.*, 1973. 2

[10] Daniel Fried, Jacob Andreas, and Dan Klein. Unified pragmatic models for generating and following instructions. *CoRR*, abs/1711.04987, 2017. 8

[11] Edward Gibson, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T. Piantadosi, and Bevil R. Conway. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790, 2017. 2

[12] Noah D. Goodman and Michael C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818 – 829, 2016. 6, 8

[13] Caroline Graf, Judith Degen, Robert X. D. Hawkins, and Noah D. Goodman. Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 2016. 8

[14] H. P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics*, pages 43–58. Academic Press, New York, 1975. 8

[15] Robert X. D. Hawkins. Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, 47(4):966–976, 2015. 2

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 8

[18] Sahar Kazemzadeh, Vicente Ordonez, Matten Mark, and Berg L. Tamara. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2, 8

[19] Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, 2015. 2

[20] Robert M. Krauss and Sidney Weinheimer. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1964. 8

[21] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. *CoRR*, abs/1804.03984, 2018. 8

[22] David Lewis. *Convention: A philosophical study*. Harvard University Press, 1969. 8

[23] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. *CVPR*, 2018. 2, 8

[24] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. 2, 8

[25] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Murphy Kevin. Generation and comprehension of unambiguous object descriptions. *CoRR*, abs/1511.02283, 2016. 8

[26] Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. Colors in context: A pragmatic neural model for grounded language understanding. *CoRR*, abs/1703.10186, 2017. 3, 8

[27] K. Varun Nagaraja, I. Vlad Morariu, and Davis S. Larry. Modeling context between objects for referring expression understanding. *ECCV*, 2016. 2, 8

[28] Maike Paetzel, David Nicolas Racca, and David DeVault. A multimodal corpus of rapid dialogue games. In *Language Resources and Evaluation Conference (LREC)*, 2014. 8

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 5

[30] Seymour Rosenberg and Bertram D. Cohen. Speakers' and listeners' processes in a word-communication task. *Science*, 1964. 8

[31] Sheng Shen and Hung Lee. Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. *CoRR*, abs/1604.00077, 2016. 4

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 8

[33] Jong-Chyi Su, Chenyun Wu, Huaizu Jiang, and Subhransu Maji. Reasoning about fine-grained attribute phrases using reference games. *CoRR*, abs/1708.08874, 2017. 8

[34] Kees van Deemter. *Computational models of referring: a study in cognitive science*. MIT Press, 2016. 8

[35] Ramakrishna Vedanta, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. *CoRR*, abs/1701.02870, 2017. 8

[36] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2017. 8

[37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2015. 8

[38] Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*, 1989. 5

[39] Ludwig Wittgenstein. *Philosophical investigations*. Macmillan, 1953. 8

[40] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. 6

[41] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2016. 8

[42] Li Yi, Hao Su, Xingwen Guo, and Leonidas J. Guibas. Syncspeccnn: Synchronized spectral CNN for 3d shape segmentation. *CoRR*, abs/1612.00606, 2016. 4

[43] Licheng Yu, Zhe Lin, Xiaohui Shen, Yangm Jimei, Xin Lu, Mohit Bansal, and L. Tamara Berg. Mattnet: Modular attention network for referring expression comprehension. *CVPR*, 2018. 2, 8

[44] Licheng Yu, Patrick Poirson, Shan Yang, C. Alexander Berg, and L. Tamara Berg. Modeling context in referring expressions. *ECCV*, 2016. 2, 8

[45] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. A joint speaker-listener-reinforcer model for referring expressions. *CoRR*, abs/1612.09542, 2017. 8

[46] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014. 8

[47] Long Zhu, Yuanhao Chen, Alan Yuille, and William Freeman. Latent hierarchical structural learning for object detection. *CVPR*, 2010. 2