# Balancing economy and utility in semantic categorisation

**Anonymous authors**
Anonymous affiliations

## Abstract

When it comes to semantic categorisation, “carving nature at is joints” is more complex than the allusion suggests. Semantic joints may not always be easily discernible, and subdivision along continuous dimensions where there are no joints may also be useful for communicative purposes. We examine these issues in an experimental setting in which participants are required to communicate about two perceptual object classes using three category labels. By manipulating the observed utility of object-label mappings, we ask whether the categories that emerge out of this kind of communicative process are shaped by such information. Inspired by information theory and rate distortion theory in particular, we conceptualise the location of category boundaries as the outcome of an iterative algorithm whereby speakers and listeners evolve a stable balance between economy of representation and utility of communication. Our results suggest that people are sensitive to the shape of an explicit utility function governing the outcome of imprecise communication as well as the representational utility achieved through categories that partition the underlying similarity space in a coherent way.

**Keywords:** categorisation; utility; communication; rational inference.

## Introduction

Two colleagues are sailing to Toronto on a zero-emissions yacht when they are shipwrecked by unseasonal weather. Each washes ashore on a different atoll, so they can only communicate via smoke signals. One important topic of conversation is the native jubjub berries. There are a variety of hard ones (good only for throwing at rats) as well as a variety of softer ones (which taste great, especially when paired with the right kind of rat). Given the limited coding scheme permitted by smoke signals, how should they decide to allocate signals to berries? Should they use one signal for hard and one for soft, or should they use more signals for the softer berries since they are more important to distinguish?

As this scenario hints, carving up a semantic space is not always as simple as segmenting along clear divides. While there is a considerable body of literature which covers how people learn lexical concepts and use them to solve inductive problems (Kruschke, 2008), relatively less is known about what factors shape the nature of these categories in the first place. Recent research suggests one possibility: that semantic and linguistic systems reflect a trade-off between simplicity and expressiveness in a variety of domains (Kemp, Xu, & Regier, 2018). For instance, as Zaslavsky, Kemp, Regier, and Tishby (2018) demonstrate, most colour-naming systems across the world’s languages exist along an optimal curve that trades off between these factors. Cross-linguistic variation occurs primarily because different lexical systems solve the trade-off in different ways.

But what factors determine exactly *how* any given lexical system solves this trade-off and where along the curve it ends up? For example, given a fixed number of category labels to describe objects in a semantic space, what determines which objects get grouped together by the same label? Are emergent systems shaped by the fact that some things might be more important to distinguish (or have higher communicative utility) than others? Or do other factors dominate, such as the similarity relations encoded in people’s mental representations? These questions are the focus of the present study.

## Evolving categories through communication

We can formalise the nature of the trade-off above as follows. Consider our two castaways: since they wish to allocate fewer category labels (smoke signals) than there are objects (types of berry), we can view their problem as one of designing a lossy compression code. Shannon (1959), in developing *rate distortion theory*, showed that there is a fundamental trade-off between the economy of representation (the level of compression achieved) and the fidelity with which the original data can be reconstructed from the compressed representation. An iterative technique for finding optimal trade-offs is known as the Blahut-Arimoto algorithm (Arimoto, 1972; Blahut, 1972). It can be applied to the design of a categorisation scheme (a mapping of labels to objects) by either using it to minimise loss (maximise utility) given a fixed number of labels or to minimise the number of labels required to keep the loss under a fixed threshold.

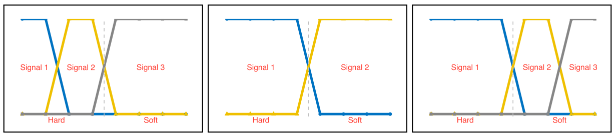
Each iteration of the algorithm (at step $t$) involves improving the encoding scheme as follows:

$$P_t(l\,|\,x) \propto \frac{P_{t-1}(l)}{\exp(\beta d(x,l))} \tag{1}$$

where $P_t(l\,|\,x)$ denotes the probability of encoding object $x$ using label $l$. In the denominator, $d(x,l)$ represents the loss incurred when object $x$ is represented via label $l$. The $\beta$ parameter controls the degree to which minimising distortion is favoured over maximising compression such that higher $\beta$ means less compression. Given a constant value of $\beta$, and holding all else equal, one label represents a better (more likely) encoding than another if it leads to less distortion. The numerator represents the frequency with which the label $l$ is used: that is,

$$P_t(l) = \sum_{x \in \mathcal{X}} P(x) P_t(l\,|\,x) \tag{2}$$

where $\mathcal{X}$ reflects the set of objects considered, and $P(x)$ is the prior probability that object $x$ will be the subject of communication. Holding the denominator equal in Eqn. 1 shows that one label represents a better encoding for $x$ than another if it is already used more frequently. Thus, starting with an arbitrary encoding, successive iterations of the algorithm fine-tune the utility and economy of the encoding scheme until no better balance can be found.



(a) (b) (c)

Figure 1: **Utility-based classification schemes**. Three alternative schemes for categorising jubjub berries, where distinguishing between hard berries (which cannot be eaten) is less useful than distinguishing between soft ones (which can be eaten with different things). Scheme (a) is non-optimal since it uses an extra (third) label but does not use it to improve utility, since the label makes an unimportant distinction (amongst hard berries). Scheme (b) minimises the number of labels (two) but at the cost of not making an important distinction (amongst soft berries). Scheme (c) uses more labels but encodes that distinction. Whether (b) or (c) represents the better design depends upon the weight given to utility as well as the need for economy of representation (fewer labels).

While further modelling details are required to make precise numerical predictions – something we return to in our later analyses – the above description gives an intuitive feel for the problem faced by our castaways. Consider the three categorisation schemes shown in Figure 1. In the first scheme (Figure 1a), two labels are used for hard berries and one for soft. Assuming all varieties of hard berry are equally good for hunting rats, subdividing hard berries comes at no gain but reduces compression. Thus, Eqn. 1 would yield a higher probability for the second encoding scheme (Figure 1b) where a single label is used to represent all hard berries. Now consider the third scheme (Figure 1c), which subdivides the softer berries. Because there is something to be gained in utility in this case (each variety of berry has the perfect companion rodent), the reduction in compression involved in introducing additional labels may be more than offset by the cost of an additional label (depending on the setting of $\beta$).

Rate distortion theory would thus predict that if people are sensitive to this kind of communicative utility then one would expect category systems that evolve through communication not to look like the one in Figure 1(a). If communicative utility is weighted heavily and the need for an expressive system is thus high, one would expect further that a system with more labels that makes an important distinction given the task in the world, as in Figure 1(c), would evolve.

However, communicative utility is not the only thing that might shape the need for a more expressive system. We might also consider category goodness, which we can think of in this context as a kind of *representational* utility. People generally have a preference for categories that are coherent in the sense that they maximise within-category similarity while minimising between-category similarity, and consistently find some categories easier to learn (Rosch & Mervis, 1975). Indeed, categories that "carve nature" at her "natural joints" – that is, categories that align naturally with the similarity structure of the underlying space – are not only easier to learn and remember, they also tend to emerge naturally out of an evolutionary process (Perfors & Navarro, 2014).

The question we consider in this work is how the pressure for expressive categories is shaped by these two kinds of utility. To what extent are the categories that emerge via a conversational interaction shaped by their communicative utility? To what extent is their structure sensitive to their representational utility, insofar as they break up the similarity space in a sensible way? This question is difficult to study via observational data based on existing world languages, because it is difficult if not impossible to estimate the factors that might affect real-world utility. We therefore investigate it in an experiment in which participants play a communication game with each other. Over the course of the game, label-object mappings emerge as people aim to maximise their score, which is calculated based on different communicative utility functions in different conditions. Does the structure of the utility function shape the categorisation schema that emerges as predicted by rate distortion theory? How does it interact with the representational utility reflected in the underlying similarity structure between the objects?

## Method

As the basis for our pre-registered empirical investigation[1] we used a two person communication game where people cooperated with each other in an effort to converge on a shared set of labels for a set of objects. In the game, people alternated roles between speaker and listener over a series of trials. On each trial the speaker saw a novel object and had to communicate the object's identity by selecting one of three provided labels. On seeing the label, the listener decided which object the speaker was referring to. Participants then received a joint score which depended on how similar the listener's choice was to the target object chosen by the speaker. By manipulating the scoring system between dyads, we investigated whether people's emergent categorisation schemes were influenced by communicative utility: did they tend to converge on categories that optimised the expected point gain given their scoring system?

### Participants

Data were collected from 351 people who participated in our experiment via Amazon Mechanical Turk. Each was paid \$5.00USD for 25-35 minutes participation. As a further financial incentive, people were aware that a bonus of \$2.50USD would be paid to players with scores in the top 10%. Each person was paired randomly with another person who arrived within two minutes of them. Those whose arrival times did not overlap with anybody else ($N = 62$) were paired instead with a computer partner. For simplicity we exclude them from the present analyses, along with data from one person whose partner failed to complete the experiment. The remaining 288 participants ranged in age between 19 and 76 (median: 37 years), comprised 52% females, and were drawn predominately from the U.S. population (98%).

[1] We cannot link to the pre-registration yet because we discovered too late that the blinded version we made includes a link to a previous pilot experiment that we unfortunately forgot to blind.

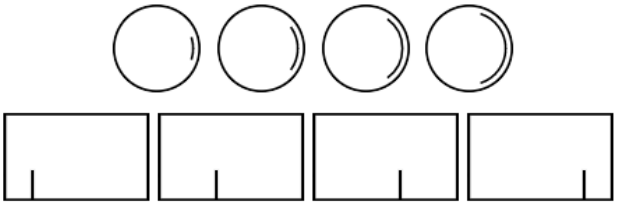


Figure 2: **Object stimuli**. The eight objects were drawn from two perceptually distinct groups that varied along one dimension within group: circles according to the central angle of the interior arc and rectangles according to the position of the short black vertical line along the bottom edge.

## Stimuli

Object stimuli were drawn from two perceptually distinct classes (circles and rectangles) and included four items of each class. As Figure 2 shows, circles varied according to the central angle subtended by an internal arc and rectangles varied in the relative position of the vertical black line along the bottom edge. As such, the stimuli defined an implicit similarity space with one main division (circles and rectangles) and continuous one-dimensional variation within each.

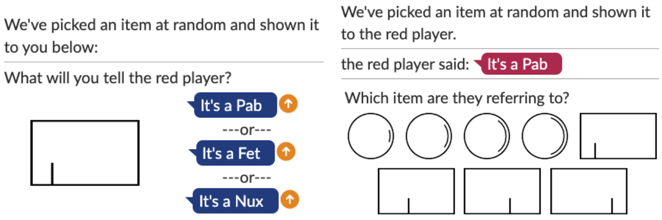
## Procedure

Our experiment was conducted in three phases. In the first, participants received independent (self-paced) training on the scoring system. They then proceeded to the main phase of the experiment where they were paired with another participant to play the communication game. Following the game, each person took part independently in a categorisation test.

**Training** An assumption of rate distortion theory, and thus a simplifying assumption in our utility-based model of communicative inference, is that the loss function (or equivalently, the utility) is known prior to communication. In the context of our experiment this meant that it was necessary for participants to be trained on the scoring system before beginning the communication game. After reading the instructions, people were shown a graphical depiction of the utility function (similar to the representation in Figure 4). Following this, they were asked eight multiple choice questions about the system; a typical question presented them with two pairs and asked them which would receive a higher score. Participants were required to repeat the training until all questions had been answered correctly.

**Communication Game** At the start of the game, participants were told that they would be playing the game (as "blue") with another player ("red") who arrived to take the experiment within two minutes of them. They were told that because they were playing with another person they would need to respond quickly on each trial, with a visible timer enforcing a two-minute time limit per response.

Each of the 40 game trials consisted of a speaker turn followed by a listener turn, with people alternating the role of speaker and listener between successive trials. As Figure 3 shows, on each trial the speaker was presented with an object drawn at random from the eight objects and asked to select one of three category labels (Pab, Fet or Nux) for the object. Upon receiving the speaker's message, the listener then selected which object they thought the speaker was referring to.



We've picked an item at random and shown it to you below:

What will you tell the red player?

It's a Pab

---or---

It's a Fet

---or---

It's a Nux

We've picked an item at random and shown it to the red player.

the red player said: It's a Pab

Which item are they referring to?

(a) Speaker turn

(b) Listener turn

Figure 3: **Game trial interactions**. On each trial, players in the communication game took turns to be (a) the speaker who selects a label to apply to the given object, and (b) the listener who is given the label and must guess which object the speaker meant to refer to.

After both speaker and listener made their selections, both players were shown the joint score they achieved on that trial. The score for every trial along with the selected objects and label were also added to a scrolling window, so that each person could see the interaction history. As a motivation for improvement, each person also saw their average score per turn along with a "good average" to aim for (85% of the optimal).

**Categorisation Test** Immediately following the communication game, each person individually took part in a test phase designed to elicit a stable snapshot of the categorisation scheme that had evolved throughout the communication game. In it, people were shown each of the eight objects one at a time in random order. This sequence was repeated three times, making 24 test trials in total. On each trial, they were asked "What would you say this is?" and given the three labels as response options. Unlike in the game, however, neither the complete set of stimuli nor the interaction history remained on-screen. This test thus required participants to rely on their memory of the categories from the game.

## Design

Our experiment manipulated Communicative utility between dyads. The utility of any given communicative interaction corresponded to the score each dyad received on that trial. The score was calculated based on the similarity between the object the speaker was asked to communicate about (the *encoded* object) and the object identified by the listener (the *decoded* object). As the utility matrices in Figure 4 show, higher points were rewarded when the two objects were identical, reflecting the fact that successful communication has higher utility. In addition, to reflect the fact that imprecise communication can also have value as long as it is not *too* inaccurate, mismatches in which the objects were sufficiently perceptually similar were also rewarded some points.

Each dyad was assigned to one of two utility conditions. In the SYMMETRIC condition ($N = 146$), both circles and rectangles have the same pattern of scores. In the ASYMMETRIC condition ($N = 142$) this is not the case: one shape yields

|  | ○ | ○ | ○ | ○ | ▭ | ▭ | ▭ | ▭ |
|---|---|---|---|---|---|---|---|---|
| ○ | 10 | 5 | . | . | . | . | . | . |
| ○ | 5 | 10 | . | . | . | . | . | . |
| ○ | . | . | 10 | 5 | . | . | . | . |
| ○ | . | . | 5 | 10 | . | . | . | . |
| ▭ | . | . | . | . | 100 | 50 | . | . |
| ▭ | . | . | . | . | 50 | 100 | . | . |
| ▭ | . | . | . | . | . | . | 100 | 50 |
| ▭ | . | . | . | . | . | . | 50 | 100 |

(a) ASYMMETRIC

|  | ○ | ○ | ○ | ○ | ▭ | ▭ | ▭ | ▭ |
|---|---|---|---|---|---|---|---|---|
| ○ | 10 | 5 | . | . | . | . | . | . |
| ○ | 5 | 10 | . | . | . | . | . | . |
| ○ | . | . | 10 | 5 | . | . | . | . |
| ○ | . | . | 5 | 10 | . | . | . | . |
| ▭ | . | . | . | . | 10 | 5 | . | . |
| ▭ | . | . | . | . | 5 | 10 | . | . |
| ▭ | . | . | . | . | . | . | 10 | 5 |
| ▭ | . | . | . | . | . | . | 5 | 10 |

(b) SYMMETRIC

Figure 4: **Communicative utility**. The score received by participants during each turn of the communication game is determined by the object labelled by the speaker (column) and the object chosen by the listener (row). When the objects match (diagonal), scores are highest. Some perceptually similar items may also earn points, albeit fewer (off-diagonal numbers), while most mismatches earn no points at all (dots). (a) In the ASYMMETRIC utility condition the values in the lower right quadrant are a 10× multiple of those in the upper left quadrant. This reflects the greater communicative utility of rectangles: getting them correct results in a higher score. The sample utility matrices shown here correspond to a situation in which the circles are the low-utility items; for half of the dyads the rectangles were the low-utility ones. (b) By contrast, in the SYMMETRIC condition the utility is equivalent for both circles and rectangles.

higher scores (i.e., higher communicative utility) than the other. (Which shape was higher, as well as the order that the shapes appeared onscreen, varied randomly between dyads). As the modelling work below demonstrates, if people are sensitive to scoring utility then one would expect their category systems to distribute two category labels within the higher-scoring shape and only one within the other, as that will maximise communication accuracy for the higher-scoring items and thus the overall score.

## Results

Our work is focused on understanding how communicative and representational utility influence semantic categorisation in language. By experimentally manipulating the scores associated with communication about different objects, we can ask whether any asymmetry in communicative utility is reflected in the nature of the categorisation scheme that evolves.

To investigate this question we analysed people's categorisation scheme on the basis of their responses during the final categorisation test, which should represent a stable snapshot of the scheme that evolved during the game. In order to allow individual schemes to be compared and considered in the aggregate, it was first necessary to deal with the non-identifiability inherent in the use of three arbitrary category labels. To this end, we applied a simple relabelling heuristic to the data, producing a mapping for each participant from the labels used in the experiment: Pab, Fet and Nux to $C_1$, $C_2$ and $C_3$. The details of this relabelling are as follows. Reflecting the structure inherent in our utility function, we first collapsed objects into related equivalence classes (with object 1 grouped with object 2, object 3 with object 4, and so on). With minimal loss of generality we identify the label produced most frequently in response to the last pair of objects and class this $C_3$. We repeat the procedure for the first pair, identifying it as $C_1$. The remaining label we denote $C_2$.[2]

After relabelling each person's responses from the categorisation phase in this way, we then obtained the aggregate categorisation curves shown in Figure 5. The plots reveal an effect of our experimental manipulation of communicative utility on the categorisation schemes that people come to adopt. In the ASYMMETRIC condition, one group of objects (rectangles or circles, depending on random assignment) has greater utility than the other. This is reflected in the fact that the second label ($C_2$, the yellow line) is not symmetrically distributed around the central vertical axis (Figure 5a). Consistent with the communicative-utility-based model predictions below, the second label was more often applied to the high-scoring objects than the low-scoring ones. In the SYMMETRIC condition however, where all objects have the same overall utility, the distribution of labels is more symmetric (Figure 5b).

To quantify the strength of the categorisation bias, we calculated the proportion of labels of each type assigned to the two different shapes: stimuli 1–4 on the left (corresponding to low-utility items in the ASYMMETRIC condition) and stimuli 5–8 on the right. Figure 5c, which plots these proportions, indicates that the emergent category systems were shaped by communicative utility. When stimuli 5–8 had higher utility (in the ASYMMETRIC condition) people were more likely to use the second category label ($C_2$, yellow) for them; when they did not (in the SYMMETRIC condition) people were not.

To quantify the strength of evidence for these findings we performed a Bayesian logistic regression, comparing two models of categorisation responses. The first model included two predictors: perceptual grouping (left or right) and stimulus order (rectangles or circles first), while the second added an interaction term between the communicative utility condition (ASYMMETRIC or SYMMETRIC) and the perceptual grouping. We found strong evidence in favour of the model with utility as a predictor ($\mathrm{BF}_{21} > 10^3$).[3]

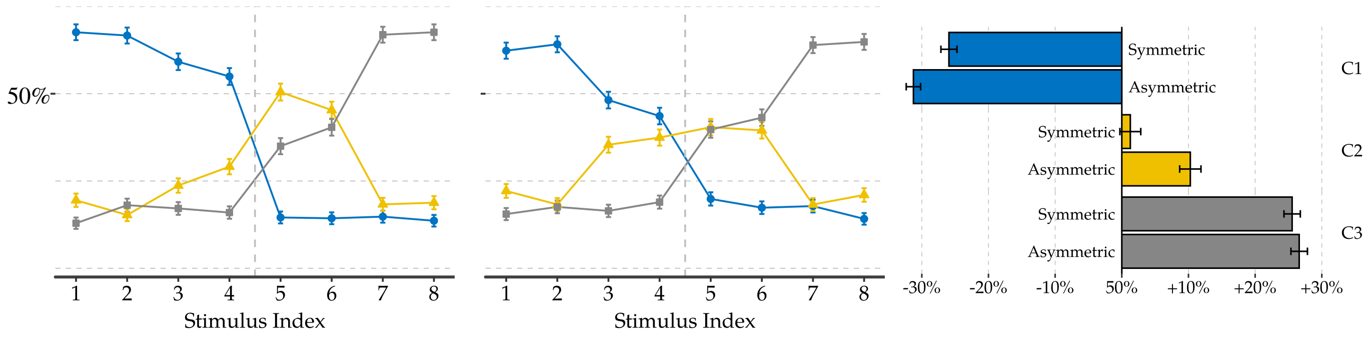
## Towards a computational model

By varying the communicative utility between conditions, our experiment allowed us to test whether it affected the shape of the emergent lexicon. Our results suggest that people do take such utility into account, albeit to a limited extent. Only when we induced asymmetry in utility did we find a corresponding asymmetry in the categorisation schemes that evolved.

But what of the strength of our finding? If people do tend to evolve systems of category labels in keeping with the kinds of optimal and stable equilibria that rate distortion theory and

[2] We tested a number of similar relabelling schemes, with different assignment orders (from left to right, right to left, left-right-middle, right-left-middle) and different granularities (1, 2 and four objects per group). The scheme used in the present analyses (object pairs with right-left-middle ordering) best fit the data using the logistic regression model discussed in the main text.

[3] Models included a random intercept for the dyad to which each participant was assigned, and were fit using default priors via the `brms` package (version 2.10.0) in R (version 3.6.1).



(a) ASYMMETRIC condition

(b) SYMMETRIC condition

(c) Categorisation bias

Figure 5: **Categorisation bias**. Evolved category structure as reflected in the labels applied by participants during the categorisation test. The $x$-axis shows the eight stimuli, arranged so that the lower-scoring objects in the ASYMMETRIC condition correspond to stimuli 1-4 and the higher-scoring ones correspond to stimuli 5-8. Plots (a) and (b) show the aggregate proportion of each label for each object (blue lines for $C_1$, yellow for $C_2$, and grey for $C_3$). In the ASYMMETRIC condition (a), the second label $C_2$, is more frequently applied to the higher-scoring objects, revealing a bias for making finer grained distinctions when communicating about high utility items. In contrast, in the SYMMETRIC condition (b) where there the utility is uniform between circles and rectangles, there is correspondingly no bias in which category the label $C_2$ applies to. (c) Bars show the proportion of each label assigned to stimuli on each side of perceptual space, with stimuli 1–4 on the left and 5–8 on the right. Bars that extend further from the central line thus indicate a stronger bias. While the process of standardising category labels determines the direction of the biases for categories $C_1$ (blue) and $C_3$ (grey), category $C_2$ (yellow) is unconstrained and thus has the scope to reflect a genuine response bias. Indeed, when there is no bias inherent in the underlying utility (SYMMETRIC), the category label is used roughly equivalently for both "sides" of perceptual space. Conversely, when a bias is present in the utility function (ASYMMETRIC), the category label is used more for the higher-scoring stimuli.

the Blahut-Arimoto algorithm predict, then we might expect a stronger effect – for example, we might expect the results from our ASYMMETRIC condition to more closely resemble the "clean" separation of categories displayed in Figure 1c. To check these intuitions it is instructive to specify a computational model of our experimental task with which more precise numerical predictions can be made.

We take the Blahut-Arimoto algorithm as the basis for our model, and assume that the speaker selects labels according to Equations 1 and 2. We adopt this algorithm because it is capable of finding the kinds of optimal trade-offs we are interested in, and because it suggests intriguing parallels with existing rational models of communicative inference (e.g., Shafto, Goodman, & Griffiths, 2014).

In order to specify how the listener decodes the speaker's label (that is, how they guess an object $\hat{x}$ given a label $l$) we assume that listeners are Bayesian reasoners who first infer a probability distribution over alternative hypotheses concerning the speaker's intended referent. We consider two ways that listeners might proceed from belief to guess. Under Probability Matching, the probability that the listener makes guess $\hat{x}$ upon seeing the label $l$ is given by:

$$P(\hat{x}\,|\,l) \propto P(l\,|\,\hat{x})P(\hat{x}). \tag{3}$$

where $P(l\,|\,\hat{x})$ follows Equation 1. $P(\hat{x})$ reflects the listener's prior belief that $\hat{x}$ is what the speaker was referring to, and is set to the uniform prior used in the game. As an alternative, listeners instead might perform Utility Matching:

$$P(\hat{x}\,|\,l) \propto \sum_{x\in X} P(l\,|\,x)P(x)u(x,\hat{x}) \tag{4}$$

where $u(x,\hat{x})$ reflects the utility of guessing $\hat{x}$ when the true referent is $x$. Using this strategy, a listener who believes $x = x_i$ more strongly than $x = x_j$ may yet prefer to guess the latter if it brings the greater expectation of utility, regardless of what they think the speaker's referent was.

Having specified how the listener makes guesses, we can now define the distortion function that the speaker and listener together work to minimise:

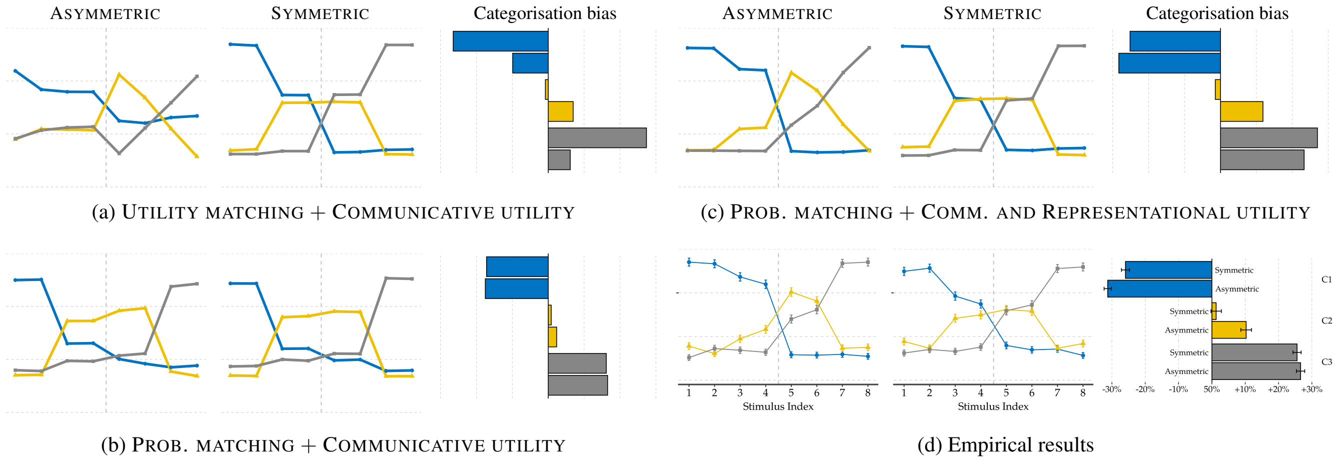
$$d(x,l) = -\sum_{\hat{x}\in X} p(\hat{x}\,|\,l)u(x,\hat{x}). \tag{5}$$

Figure 6 plots categorisation curves aggregated over 1,000 simulated games for three alternative configurations of the model. In all plots the stable equilibria found by the iterative technique depend upon the starting point, which means that the aggregate curves, much like our empirical results, represent a range of solutions. What differs across the three models are the frequency with which different solutions occur.

The model variant in Figure 6a simulates the Utility Matching listener, based on the communicative utility defined in the game. Because both the speaker and listener take utility into account, this variant is the most aggressive at attempting to maximise utility. As a result, much of the time it finds the (non-intuitive) global optimum[4] whereas people do not. Largely for this reason it fails to capture human performance as well as the other models.

The second model (Figure 6b) simulates dyads who work to minimise the distortion function that includes utility (Equation 5) but assumes that the listener guesses the referent of the speaker based solely on likelihood (Equation 3). As a result of this, sub-optimal solutions where two labels are dedicated to the low value objects nonetheless represent stable

[4] All three labels are effectively associated with the high utility objects but generated at random in response to low value objects.



(a) UTILITY MATCHING + COMMUNICATIVE UTILITY

(c) PROB. MATCHING + COMM. AND REPRESENTATIONAL UTILITY

(b) PROB. MATCHING + COMMUNICATIVE UTILITY

(d) Empirical results

Figure 6: **Model simulations**. Evolved category structure reflecting the labels generated by 1,000 simulated dyads (speaker and listener) playing the communication game. Three variants of our computational model based on the Blahut-Arimoto algorithm are used. All variants assume a speaker who selects labels according to Eqns. 1 and 2, and a rational Bayesian listener. In (a) and (b), the model reproduces the utility matrices as used in the game. However, while in (a) the listeners use "utility matching" to select among possible interpretations of the label, the listeners in (b) use "probability matching" instead. The listeners and speakers in (c) behave as in (b) except that they use a utility matrix which accounts for perceptual similarity as well. Model (c) appears to match human performance most closely.

equilibria under the model. The similarity in categorisation bias between ASYMMETRIC and SYMMETRIC utility evident in Figure 6b reflects the relatively low frequency with which the utility-biased solution is found.

Both models so far reflect *communicative* utility but may not sufficiently capture *representational* utility, since the perceptual similarity between items in each quadrant is not captured by the communicative utility matrices in Figure 4. To explore the value of representational utility further, the third model (Figure 6c) takes the second model as a starting point but modifies the upper left and lower right quadrants of the utility matrices as follows:

$$k\begin{bmatrix} 10 & 5 & 0 & 0 \\ 5 & 10 & 0 & 0 \\ 0 & 0 & 10 & 5 \\ 0 & 0 & 5 & 10 \end{bmatrix} \longrightarrow k\begin{bmatrix} 10 & 5 & \mathbf{2} & \mathbf{1} \\ 5 & 10 & \mathbf{5} & \mathbf{2} \\ \mathbf{2} & \mathbf{5} & 10 & 5 \\ \mathbf{1} & \mathbf{2} & 5 & 10 \end{bmatrix}$$

where $k = 1$ or 10. Consequently, the model avoids solutions where a given label is split between circles and rectangles. As a result, the match between model and human performance is significantly improved.

These findings suggest that listeners are sensitive to referential accuracy, and prefer to match speakers' intended referents rather than maximise utility alone. They also suggest that people take both communicative and representational utility into account when evolving category systems. The challenge for us in future work is to establish principled models of why people's utility functions take the shape they do in the first place. In particular, more work is required to fully understand where what we are calling "communicative utility" and "representational utility" come from. Do people's representations of conceptual similarity and generalisability amongst objects already take into account the need for semantic precision in communication? Or did this emerge in our experiment because people were placed into an explicitly communicative context?

## Conclusion

> *"I trust in nature for the stable laws of beauty and utility. Spring shall plant and autumn garner to the end of time."*
> – Robert Browning.

Nature, it seems, has a knack for evolving the kinds of exquisitely poised equilibria to which Browning alludes. Our results join an emerging body of evidence that examines how analogous forces may be at play in the evolution of language, bringing language embeddings of real world concepts into correspondence with both the mental representations of those concepts and the utility of communicating about them.

## References

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, *18*(1), 14–20.

Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, *18*(4), 460–473.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Ann. Rev. Ling.*, *4*, 109–128.

Kruschke, J. (2008). Models of categorisation. In R. Sun (Ed.), *The cambridge handbook of computational psychology* (pp. 267–301). Cambridge University Press.

Perfors, A., & Navarro, D. (2014). Language evolution can be shaped by the structure of the world. *Cognitive Science*, *38*(4), 775–793.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.*, *7*(4), 142–163.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *31*(115), 7937–7942.