Communication Leads to the Emergence of Sub-optimal Category Structures

Catriona Silvey (C.A.Silvey@sms.ed.ac.uk), **Simon Kirby**, **Kenny Smith**
Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences,
University of Edinburgh, Dugald Stewart Building,
3 Charles Street, Edinburgh, EH8 9AD, UK

Abstract

Words divide the world into labeled categories. Languages vary in the categories they label, sometimes to the point of making cross-cutting divisions of the same space. Previous work suggests two opposing hypotheses about how communication contributes to category emergence: 1) these spaces lack an objective shared similarity structure, and communication dynamically creates one of a number of optimally shareable category structures; 2) the category structures resulting from communication are not necessarily optimal, but diverge from a shared similarity space in language-specific ways. We had participants categorize images drawn from a continuous space in two conditions: a) non-communicative, by similarity, b) communicative, dynamically creating categories when playing a partnered communication game. The memory demands of communication lead to reliance on salient images and early conventions, resulting in non-optimal category structures compared to non-communicative participants. This supports the hypothesis that communication leads to categories that diverge non-optimally from a shared similarity space.

Keywords: communication; category structure; category emergence; language evolution

Introduction

Words divide the world into labeled categories. Languages vary in the categories they label, with some languages making coarser, finer, or even cross-cutting distinctions relative to how other languages carve up the same space (Bowerman & Choi, 2001; Malt, Sloman, & Gennari, 2003). Work is ongoing to quantify and classify this variation (Majid, Jordan, & Dunn, in progress). The mechanism by which a set of labeled categories emerges in a given language is however unclear. One hypothesis is that at least for some domains (e.g. spatial relations, containers), there is no one perceptually obvious way to divide the space into categories: there are several potential ways an individual observer could draw category boundaries (Bowerman, 2000). Some researchers have built on this idea to suggest that the process of communication itself structures a previously unstructured space, making categories that are optimally shareable between communicators (Freyd, 1983; Markman & Makin, 1998; Steels & Belpaeme, 2005; Voiklis & Corter, 2012). However, cross-linguistic work by Barbara Malt and colleagues on similarity perception versus labeling shows that, while the labeled categories of different languages do indeed diverge from each other, speakers of different languages still perceive the similarities between the objects in comparable ways (Malt, Sloman, Gennari, Shi, & Wang, 1999). This suggests that the categorization systems of different languages can in fact superimpose a range of divergent structures on a space that has a shared underlying similarity structure. These two accounts suggest radically different roles for communication in the emergence of categories.

The current experiment contributes to this debate by investigating how humans categorize a set of images designed to have unclear category boundaries. The participants categorize the images in one of two conditions: a non-communicative condition, where solo participants divide the images into categories according to similarity, and a communicative condition, where pairs of participants play a communication game with the images. The results shed light on the effect of communication on category structure, suggesting that the categories created by communication can and do diverge from a relatively shared similarity space, even in a stimulus set designed to have ambiguous boundaries.

Method

Participants were assigned to two conditions. In the non-communicative condition, participants divided a continuous space of images into labeled categories on the basis of similarity. In the communicative condition, pairs of participants played a communication game using the same continuous space of images. Participants in this condition produced labeled categories via the words they used to communicate each target image in the last two rounds of the experiment. The category systems the participants produced in the two conditions were then compared.

Stimuli

The set of images used in the experiment is shown in Figure 1. The four corner images were generated using PsychoPy software (Peirce, 2007). For each image, a random number generator assigned x and y positions for the five vertices, and the resulting shape was drawn. Morphs between these images were then generated by shifting the vertices towards each of the corners, according to a weight defined by inverse Euclidean distance (Matthews, 2009), to create a total set of 25 images. The 'objective' Euclidean distance between the images in the space may of course not correspond to perceptual similarity (see, e.g., Smith & Heise, 1992); however, in pilot experiments, participants showed variation in where they drew the category boundaries, making these stimuli suitable for the current study.

Labels

To control for any effects on participants' categorizations arising purely from the use of labels (Lupyan, Rakison, & McClelland, 2007), words to label the categories were provided in both the non-communicative and communicative conditions. Lists of 25 CVCV nonsense words were generated by combining consonants and vowels randomly selected

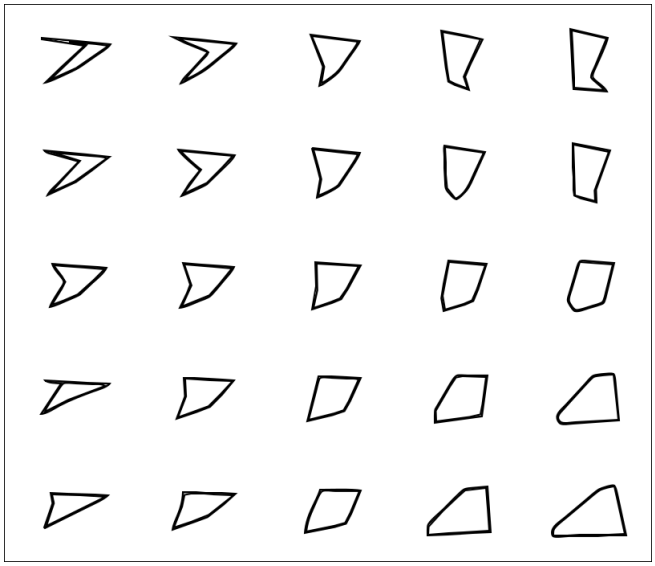

Figure 1: The stimuli used in the study (lines thickened for clarity).

from the whole alphabet (e.g., *zipi*, *gisa*, *wada*). Since we expected that participants would use known crossmodal associations between attributes of words and attributes of the images in assigning category labels (e.g. voiceless stops and spikiness, Nielsen & Rendall, 2011), we assigned the same wordlist to a yoked triple of two non-communicative participants and one communicative pair, so that in the analyses, any peculiar effects of a particular wordlist would apply equally across the conditions.

Participants

Participants were 42 students at the University of Edinburgh (30 female, median age 23). 20 took part in the non-communicative condition. The non-communicative experiment took 15 minutes. Participants were paid £2. 22 participants (randomly assigned into 11 pairs) took part in the communicative condition. The communicative experiment took an hour. Participants were paid £7, and each member of the pair with the highest communication score was awarded a £10 Amazon voucher. One pair failed to complete the experiment within an hour and so was excluded from analyses.

Procedure

Non-Communicative Condition Participants were presented with a randomized onscreen array of all 25 images and a set of words to label categories. To avoid cueing the participants to produce a particular number of categories, only one word was initially shown on screen: participants could reveal new words at any time, and were told that a) they could use as few or as many words as they wanted, and b) they did not have to use all the words they had revealed. Participants could reveal a new word at any stage, up to 25 words. They were instructed to label similar images with the same word and different images with different words.

Communicative Condition Participants communicated via computer terminals in separate cubicles. In a communication trial, one participant was assigned as the sender and one as the receiver. The sender was presented with a randomized onscreen array of all 25 images, one of which was selected with a red box to indicate it was the target. The sender was also presented with one initial word. The sender could reveal a new word at any stage, up to 25 words. Any words they had revealed on a previous trial remained visible on their screen for all subsequent trials. The participant was instructed to choose a word that would help the receiver pick out the target from the array of images.

Once the sender had picked a word, the receiver was presented with a randomized onscreen array of all 25 images and the word the sender had chosen. The receiver was instructed to select the image the sender had wanted to communicate.

Once the receiver selected an image, both participants were presented with a feedback screen. The feedback screen showed the word the sender had used, the target image, the image the receiver had selected, the score for the trial, and the running score for the whole experiment. The score for each trial was calculated on the basis of the inverse Euclidean distance between the target and the image the receiver selected, from a minimum of 1 up to a maximum of 15 (for correctly picking the target).

After each communication trial the sender and the receiver swapped roles. The experiment consisted of 100 communication trials divided into 4 rounds. Each round featured the 25 images as targets in a randomized order. The randomized lists were balanced such that each participant was the sender for every target image once in the first half of the experiment, and once in the second half.

The first two rounds of the experiment were not incorporated into the categorization analysis, as it was expected that at this stage a system would still be emerging. Participants' categories were therefore taken from the last two rounds of the experiment. Success scores were taken from the whole experiment.

Dependent Variables

Number of Categories The number of categories each participant produced was recorded.

Variation in Category Size To achieve a measure of variation in category size that took the number of categories into account (since more categories would generally contain fewer images each), the number of images in each category was divided by the expected number of images in each category, if images were distributed equally. For example, if a participant had 5 categories, an equal distribution would be to place 5 images in each category: if one of their categories in fact had 10 images, this would produce a value for that category of $10/5 = 2$. The range of these values was then taken as a measure of variation in category size adjusted for the number of categories (with a minimum value of 0 in the case of perfectly balanced categories).

Category Alignment Two measures were taken to compare participants' categories and quantify their alignment. The first, the Rand index (Rand, 1971), consists of a pairwise comparison of whether participants tended to place images in the same category or different categories. The calculation produces a value bounded from 0 to 1, where 1 is perfect alignment. The second, V-Measure (Rosenberg & Hirschberg, 2007), is based on variation of information between the groupings, normalized to compensate for differences in number of categories. This measure also ranges from 0 to 1 where 1 is perfect alignment. Two further measures, the Variation of Information measure on which V-Measure is based (Meilă, 2003) and an adjusted version of Cramer's phi (Wills & Mclaren, 1998) were considered, but were found to produce incongruent results when applied to groupings with divergent numbers of categories. Since the variable of interest was participants' categories rather than the words they used, the alignment measures were taken irrespective of whether participants used the same words: i.e., if two participants put the same set of images in a labeled category but used different labels, they would count as fully aligned for this category.

Hypotheses

For the non-communicative participants, there is no particular incentive to divide the images into more or fewer categories (beyond the minimal assumption that, in being asked to sort the images, the participants are unlikely to place them all in one category). This condition therefore functions as a baseline for assessing the variability of the participants' categorization of the images without communication. The expectation is that with no strong motivation to behave in any particular way, participants' categorization performance will vary.

For the communicative participants, the pressures on their emergent categorization systems are more complex. The only way to attain a perfect communication score with this stimulus space and scoring system is to have a unique label for each image, i.e. 25 words in total, with 25 corresponding categories containing one image each. However, participants' memory constraints will likely prevent this from happening in the experiment. More generally, then, for a given number of words, the optimal strategy is to apply each word to an equal number of images in the space, in a contiguous region (Gärdenfors, 2000). Participants who converge on a system like this would maximize their possible score across all rounds of communication. Figure 2A shows an example of this kind of optimal system. When the sender uses a word corresponding to one of the categories, the receiver can adopt the strategy of picking a central member of the category, thus ensuring their response is a maximum of 1.4 Euclidean distance units (or one diagonal step) from the target. Figure 2B shows, by contrast, a non-optimal system with the same number of categories. This system is non-optimal for two reasons. 1) The number of images in each category is less balanced (one category contains only two images, while another contains ten). This means that when the sender uses the word for the bigger category, the probability of the receiver selecting an image close to the target is lower. 2) The images belonging to some categories are spread across different regions of the space and do not form contiguous regions. This raises the probability of a receiver selecting an image some distance away from the target, even if she shares this set of categories with the sender. It is worth noting that the spaces we categorize in the real world may not have this kind of smooth continuous structure, and so the regular contiguous regions of Figure 2A may be more difficult to achieve. However, in the context of this experiment, if communication does give rise to optimally structured categories, this is the kind of system we would expect to see emerging.

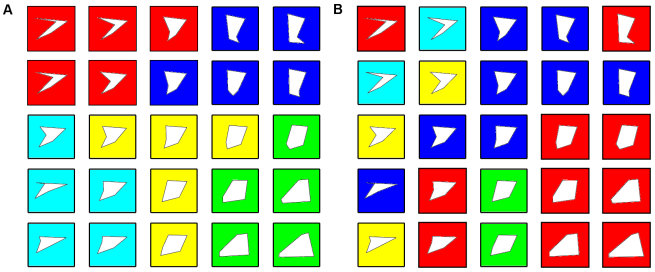

Figure 2: A) An example of a category system optimally structured for communicative success. B) A non-optimal system with the same number of categories.

Results

A linear trend ANOVA found that communicative success increased over the 4 rounds of the experiment, $F(1,9) = 18.66$, $p = .002$ (Figure 3). Participants' overall success was significantly above chance, $t(9) = 4.21$, $p = .002$.

Participants in the communicative condition used significantly more labeled categories ($M = 9.95$, $SD = 3.98$) than participants in the non-communicative condition ($M = 6$, $SD = 1.37$), Mann-Whitney $U = 60$, $z = -3.54$, $p < .001$. Communicative participants also showed significantly more variance in how many labeled categories they used, Levene's test (1,36) = 16.47, $p < .001$. Pairs who communicated together, however, showed no significant difference in the number of categories they used, $t(18) = -0.38$, $p = .7$, showing that this effect came from differences between, rather than within, communicative pairs. Thus, even though the non-communicative participants had less motivation to converge on a particular number of labeled categories, they were more consistent in the number they produced than the communicative participants.

Participants in the communicative condition also varied significantly more in the size of their categories, when number of categories was taken into account (category size variation as described in Methods $M = 1.54$, $SD = 0.35$, compared to non-communicative participants, $M = 1.17$, $SD = 0.4$). That is, images were more unevenly distributed across categories in the communicative condition, $t(38) = 3.13$, $p < .005$. Surprisingly for the communication-as-alignment hypothesis, communicative pairs' groupings did not align sig-

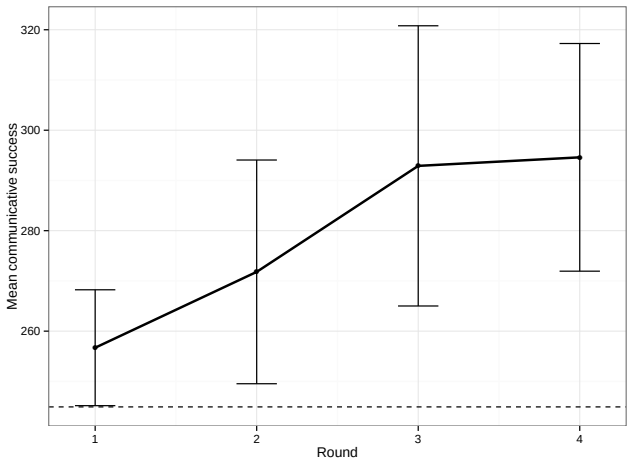

Figure 3: Average communicative success over rounds in the experiment. Dotted line shows chance. Error bars show 95% confidence intervals.

nificantly more than non-communicative participants' (by-language analysis: paired-samples t-test $0.47 < t(9) < 0.63$, $p > .5$, by-subjects analysis: independent t-test $-0.42 < t(18) < 0.63$, $p > .4$). Neither did communicative success correlate significantly with either of the alignment measures, $r < .51$, $p > .14$.

To test the hypothesis that communicative participants within a pair were more aligned than communicative participants who were not paired with each other, an analysis was run comparing the alignment scores for the true pairs with alignment scores for shuffled pairs (participant 2 paired with participant 3, etc.). A similar analysis was run for the non-communicative pairs, comparing alignment of those who shared the same wordlist with those who had different wordlists. Non-communicative participants displayed equivalent levels of alignment whether or not they used the same wordlist, $t(9) < 0.8$, $p > .58$. For communicative participants, one of the alignment measures (Rand index) tended towards being significantly higher for participants who communicated in a pair than participants who did not, $t(9) = 1.88$, $p = .093$, suggesting that communicative participants were marginally more aligned within-pair than between-pair in terms of which pairs of images they categorized together. For the second alignment measure, V-Measure, no significant difference was found, $t(9) = 1.22$, $p > .25$.

Discussion

The results are somewhat surprising for the hypothesis that communication creates optimal structure in previously variably structured spaces. Communicative participants produced categorizations that were generally non-optimal for maximizing communicative success, as defined in Hypotheses above. This is not merely a property of how humans perceive this particular space, as shown by the contrast with the non-communicative condition, where participants' categories

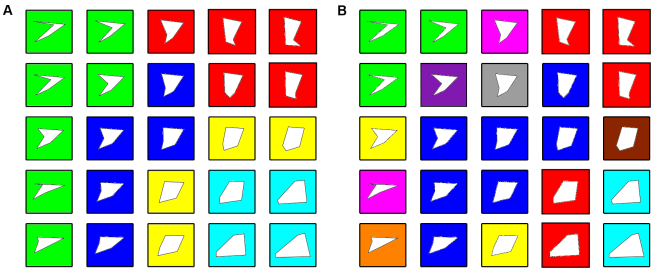

Figure 4: A) A typical non-communicative participant's categories. B) A typical communicative participant's categories.

were generally more balanced in size, carving up the space in a way that would actually be more optimal by this definition. Figure 4 shows a typical example of A) a non-communicative participant's categories and B) a communicative participant's categories. It is notable that several categories in B are also non-optimal in that they cover non-contiguous regions of the space (e.g. red and yellow categories). The heatmaps in Figures 5 and 6 show more generally how communicative participants' categories were more dispersed (Figure 6) compared to non-communicative participants, who tend to clump more around certain pairings or groups to form their categories (darker regions in Figure 5).

Why did communicative participants divide up the space so differently from non-communicative participants? As mentioned in Hypotheses above, the communicative task exerts a considerable memory demand on participants: although they are presented with the full image space on each trial, they still have to remember which word applies to which image or group of images over the course of the experiment. This exerts a pressure to create a system that is optimized not just for communicative success, but also for learnability.

Aids to learnability in this experiment might include particularly salient words, images, and pairings between them, or felicitous early successes that lead to the forming of conventions. These conventions, once established, may then prove too valuable to shift in favor of more optimally structured categories. Both of these aids to learnability (salient images/words and early successes) are mentioned by participants in the post-experiment questionnaire. Typically, when asked to draw the images they remember, participants could draw from memory two to five salient images and their associated words, but were unclear on other regions of the space. Thus the memory demands of the task, and the fact that participants have to establish a system from scratch, make the salience of individual images and early established conventions important factors determining the shape of each participant's final categorization system.

The possibility that different images in the set had differing salience is also supported by the success heatmaps in Figure 7. The heatmap in Figure 7A shows which target images led to higher success scores for participants. The pattern here is at odds with Figure 7B, which shows the relative expected

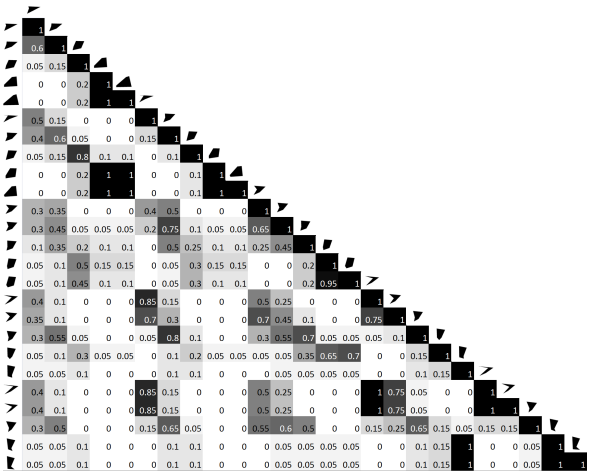

Figure 5: Heatmap visualizing how often non-communicative participants placed pairs of images in the same category. Darker areas indicate pairs more often categorized together.

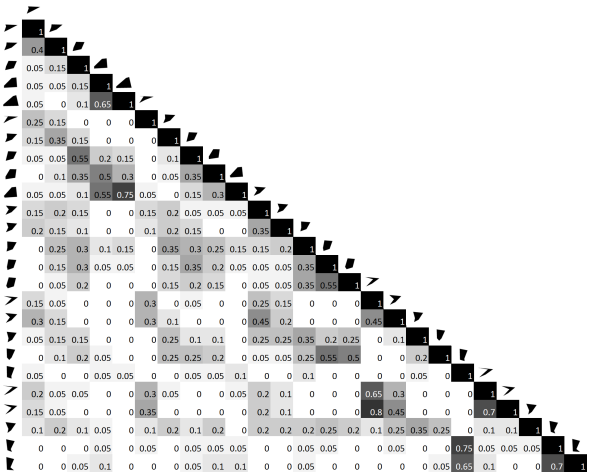

Figure 6: Heatmap visualizing how often communicative participants placed pairs of images in the same category. Darker areas indicate pairs more often categorized together.

chance levels of success for each image: images in the middle have more low-ED neighbors, so the probability of a higher score goes up when they are the target. The fact that panels A and B differ shows that participants' success with particular images is boosted by some other factor.

Panel C shows a heatmap of this boost – darker images are those whose overall communicative success rate is highest compared to what the chance-based map in panel B would predict. The likely explanation is that these images have higher salience for participants, making them act as Schelling points between sender and receiver. The striking finding that communicative success is not correlated with overall alignment could therefore be explained by participants consolidating success on a few images, leaving other areas of the space more sparsely covered.

While Figure 7 suggests that the salience of particular images may be shared across all pairs, early conventions are more likely to vary between pairs due to the randomized presentation of targets. This could explain the tendency towards higher pairwise (Rand index) alignment within pairs than between pairs, as reported in the Results. Despite the low levels of alignment overall, communicative pairs' language-specific early conventions may bring them more into agreement on how they categorize specific small groups of images.

As mentioned above, the pressures on the participants in the two conditions were substantially different: participants in the non-communicative condition interacted with the stimuli more briefly and without memory constraints, as well as lacking the pressure to create more categories imposed by the communicative task. Future work could investigate how participants divide up the space non-communicatively under the same time and memory constraints as the communicative participants, thus disentangling the effects of these constraints from the effects of communication. The non-communicative condition in this study still serves as a useful

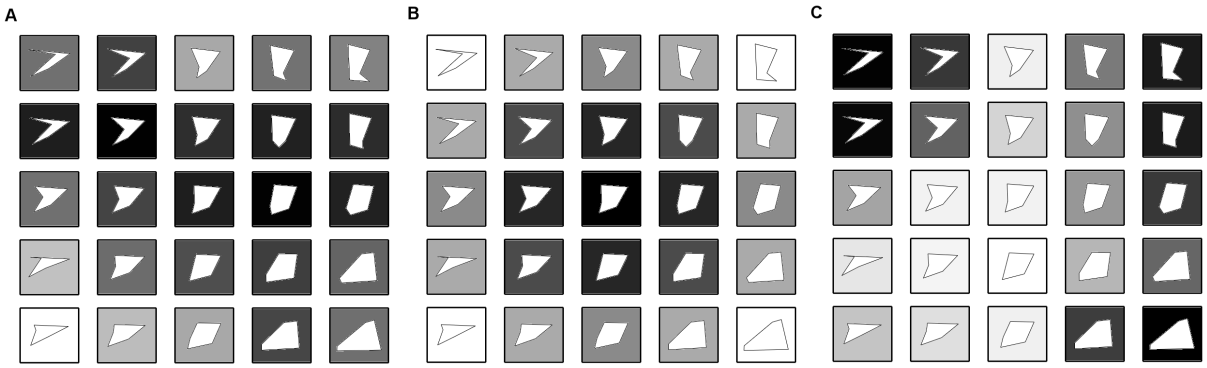

Figure 7: Heatmaps showing which target images produced higher per-round success scores. Darker images produced higher scores. A) Map of overall success per image in the experiment. B) Map of expected chance success rates per image. C) Difference between maps A and B. Darker images are those whose success rates are boosted highest beyond expected.

baseline, however, for participants' perceptually based divisions of the space.

The outcome of this study – that communication does not necessarily optimize category structures, but can create uneven and suboptimal structures compared to non-communicators' division of the same space – is reflected in our experience of real language, where words vary widely in whether they specify small regions of semantic space or broad undifferentiated regions. The existence of the latter kind of word does not necessarily mean the users of the language do not perceive the differences between sub-parts of the region it covers: only that, for reasons of salience, or constraints imposed by the history and development of conventions in the language, these internal differences lack category labels. An important additional pressure in real language, not modeled in this study, is that different regions of semantic space may also have different functional importance, motivating coarser or finer-grained distinctions in different regions. However, these results show that even in the absence of functional reasons for uneven division of a space, communication can lead to the establishment of categories that may not align with non-communicative similarity perception.

Conclusion

Communication is not a simple process of mapping words onto pre-shared perceptual categories. Even if communicating partners agree on the underlying structure of the space they are talking about, the categories that emerge from communication can diverge in surprising ways, both from the underlying similarity space and from the category structure that would be most optimal for communicative success. Constraints on learning, salience effects, and the impact of early conventions on a language's development all contribute to shaping an emergent system of labeled categories.

Acknowledgments

CS is supported by an AHRC PhD studentship. Thanks to Christos Christodoulopoulos for help with alignment measures and with networking for the communication experiment, and to Mark Atkinson and Andrea Ravignani for help with piloting.

References

Bowerman, M. (2000). Where do children's word meanings come from? Rethinking the role of cognition in early semantic development. In L. Nucci, G. Saxe, & E. Turiel (Eds.), *Culture, thought and development* (pp. 199–230). Mahwah, NJ: Lawrence Erlbaum.

Bowerman, M., & Choi, S. (2001). Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 475–511). Cambridge: Cambridge University Press.

Freyd, J. (1983). Shareability: The social psychology of epistemology. *Cognitive Science*, *7*(3), 191–210.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–83.

Majid, A., Jordan, F., & Dunn, M. (in progress). *Evolution of semantic systems.* http://www.mpi.nl/departments/other-research/research-consortia/eoss. (Online; accessed 22 April 2013)

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230–262.

Malt, B. C., Sloman, S. A., & Gennari, S. P. (2003). Universality and language specificity in object naming. *Journal of Memory and Language*, *49*(1), 20–42.

Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General*, *127*(4), 331–54.

Matthews, C. (2009). *The emergence of categorization: Language transmission in an iterated learning model using a continuous meaning space*. Unpublished master's thesis, University of Edinburgh.

Meilă, M. (2003). Comparing clusterings by the variation of information. In B. Schölkopf & M. K. Warmuth (Eds.), *Learning theory and kernel machines* (pp. 173–187). Berlin: Springer–Verlag.

Nielsen, A., & Rendall, D. (2011). The sound of round: Evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology*, *65*(2), 115–24.

Peirce, J. W. (2007). PsychoPy–psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 410–420).

Smith, L. B., & Heise, D. (1992). Perceptual similarity and conceptual structure. In B. Burns (Ed.), *Percepts, concepts and categories* (pp. 233–272). Amsterdam: Elsevier B.V.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*, 469–529.

Voiklis, J., & Corter, J. E. (2012). Conventional wisdom: Negotiating conventions of reference enhances category learning. *Cognitive Science*, *36*(4), 607–634.

Wills, A. J., & Mclaren, I. P. L. (1998). Perceptual learning and free classification. *The Quarterly Journal of Experimental Psychology*, *51B*(3), 235–270.