UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Semantic uncertainty guides the extension of conventions to new referents

Permalink

https://escholarship.org/uc/item/2hx7k0gw

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors

Eliav, Ron
Ji, Anya
Artzi, Yoav
et al.

Publication Date

2023

Peer reviewed

Semantic uncertainty guides the extension of conventions to new referents

Ron Eliav[1], Anya Ji[2], Yoav Artzi[2], Robert D. Hawkins[3*]

[1] Department of Computer Science, Bar-Ilan University
[2]Department of Computer Science, Cornell University
[3]Princeton Neuroscience Institute, Princeton University
[*]Correspondence to `rdhawkins@princeton.edu`

Abstract

A long tradition of studies in psycholinguistics has examined the formation and generalization of *ad hoc* conventions in reference games, showing how newly acquired conventions for a given target transfer to new referential contexts. However, another axis of generalization remains understudied: how do conventions formed for one target transfer to *completely distinct targets*, when specific lexical choices are unlikely to repeat? This paper presents two dyadic studies ($N = 240$) that address this axis of generalization, focusing on the role of nameability — the *a priori* likelihood that two individuals will share the same label. We leverage the recently-released KILOGRAM dataset, a collection of abstract tangram images that is orders of magnitude larger than previously available, exhibiting high diversity of properties like nameability. Our first study asks how nameability shapes convention formation, while the second asks how new conventions generalize to entirely new targets of reference. Our results raise new questions about how *ad hoc* conventions extend beyond target-specific re-use of specific lexical choices.

Keywords: convention; communication; learning; generalization; abstraction

Introduction

A core problem that all theories of reference must address is the problem of *generalization*. When faced with a novel referent, we must somehow extend or combine existing items from our lexicon to produce (as a speaker) or interpret (as a listener) an utterance with the new meaning. This problem was sidestepped under classical truth-conditional approaches to reference, where the complete set of referents it applied was built into its literal meaning (Dale & Reiter, 1995; Lewis, 1970). More recently, however, a family of probabilistic approaches to reference have emerged to handle graded phenomena like vagueness (Goodman & Lassiter, 2015) and typicality (Degen et al., 2020). Under these accounts, interlocutors maintain graded *uncertainty* about exactly what an expression will mean to others (Bergen et al., 2016; Potts et al., 2015). Hence, generalization is inherently risky. There is no guarantee that a given referential meaning will be shared by one's partner, and the reference may fail.

How, then, do interlocutors decide which expressions would be more or less effective when faced with something new? One important source of evidence comes from a classic line of work using repeated reference games with deliberately ambiguous referential targets like abstract drawings or tangram shapes (Clark & Wilkes-Gibbs, 1986; Hawkins et al., 2020; Krauss & Weinheimer, 1964). These studies reliably
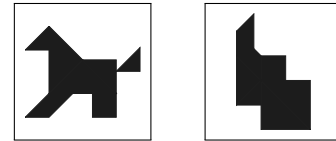

Figure 1: Example tangram images from the KILOGRAM dataset. The one on the left has relatively high nameability, and the one on the right lower.

establish the basic conditions for studying referential generalization: participants must rely on their pre-existing lexical resources to talk about something new. A key finding from these studies is that participants typically begin with verbose descriptions appealing to multiple features of the target ("a dog looking left with its tail pointed up"), but gradually form conventions and converge to a set of short labels ("the dog").

Classic studies have examined how newly-formed conventions transfer when the same target appears with a new *social partner* (Brown-Schmidt et al., 2015; Metzing & Brennan, 2003; Wilkes-Gibbs & Clark, 1992), or in a new *referential contexts* (Brennan & Clark, 1996; Ibarra & Tanenhaus, 2016). However, a critical axis of generalization has been comparatively underexamined: the problem of generalization to new referential *targets* (Nölle et al., 2018; Raviv et al., 2022; Shepard, 1987). If interlocutors are not simply memorizing a one-to-one mapping but achieving some form of broader *conceptual alignment* (Clark, 1996; Stolk et al., 2016), then we should observe a gradient of generalization to nearby locations in the same conceptual space. Probabilistic accounts of convention formation (e.g. Hawkins et al., 2022) naturally capture this gradient via hierarchical inference. If people update their beliefs about whether a given referring expressions will be understood for the *same target* in the future, they should also (more weakly) update their beliefs about other targets in the same distribution.

We approach this problem by considering the property of *nameability* — roughly, the *a priori* likelihood that two individuals will prefer the same label before interacting (Hupet et al., 1991; Zettersten & Lupyan, 2020). For a higher-nameability referent, such as the left image in Figure 1, most speakers will be expected to extend the same familiar label *dog*. Meanwhile, for a lower-nameability referent, such as the image on the right, different speakers are expected to at-
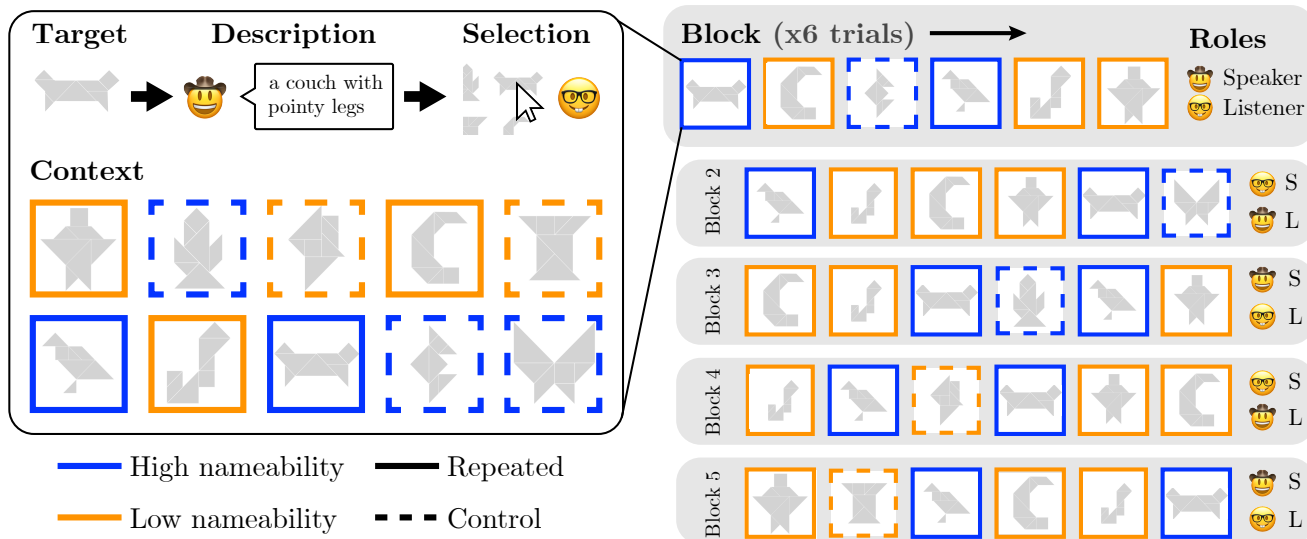
Figure 2: **Experiment 1 design**. On the left, we show an example of a single reference game trial, using a mixed context of ten tangrams (borders added for this graphic). The context remains the same for all blocks and trials. On the right, we show a full target sequence consisting of five blocks. In each block, we randomly inserted a single control tangram target (dashed border) among the repeated targets (solid border).

tempt to extend very different labels (e.g., "lizard", "robot hand", or even "lighter"). A recent study in the color domain has suggested that speakers make fairly accurate predictions about whether a given label will be shared by others (Murthy et al., 2022), although this may not be the case for all domains (Lupyan et al., 2023; Martı et al., 2023; Wang & Bi, 2021).

In this paper, we investigate the effect of nameability on referential generalization in two studies, drawing on the recently introduced KILOGRAM collection of well-normed tangram stimuli. Our studies employ the well-studied repeated reference game scenario (Clark & Wilkes-Gibbs, 1986), where two participants repeatedly communicate about a set of abstract images. On each trial, one participant, who is assigned the speaker role, describes a prespecified target image from a set of image for the other participant, who is the listener, to select. Success is measured as the listener correctly selecting the target image. The first study focuses on the effect of nameability when reference targets repeat and conventions are formed, while the second studies the transfer of such conventions to new target images. The studies advance our prior understanding of generalization in reference along multiple aspects: (a) we extend the study of repeated reference games to a much larger set of abstract tangram shapes with a higher range of nameability; (b) we evaluate the impact of nameability on the formation of conventions; and (c) we measure the impact of transfer of conventions to new reference targets within a new context.

Experiment 1: Manipulating nameability

Participants We recruited 60 pairs of participants from Prolific, based on preregistered inclusion criteria (English as first language and location based in US or UK). We excluded 8 pairs of participants, because their games contained more than 20% empty responses. Participants provided informed consent in accordance with the institutional IRB. Each game lasted an average of 23 minutes and participants were given a base payment of $4.25 (approximately $11 per hour) with a performance bonus up to $0.90.

Stimuli We designed a reference game using the black-and-white tangram shapes from the KILOGRAM dataset (Ji et al., 2022). Each shape in KILOGRAM was previously normed for nameability using a metric called Shape Naming Divergence (SND), computed over naming annotations included in the dataset. This metric is defined as the mean proportion of words in each description that do not appear in any other description for that tangram. For example, if all annotators of a tangram used the one-word description "bird", the SND would be 0 because there are no unique words; if all annotators used distinct one-word descriptions, the SND would be 1. We sorted all 1016 tangrams in the dataset by SND and selected the top 100 tangrams to use for the *low-nameability* condition and the bottom 100 for the *high-nameability* condition to ensure maximal differentiation.

For each pair of participants, we sampled a context of 10 tangrams, 5 tangrams from the high-nameability set and 5 tangrams from the low-nameability set. We ensured that tangrams in the context are sufficiently distinct to avoid challenging informativity pressures (e.g. two high-nameability tangrams that both have the consensus label *bird*). For each tangram, we extracted the head word from each of the 10 KILOGRAM descriptions using SpaCy v3 (Honnibal et al.,

2020) dependency parser. We then set a threshold of 10% overlap in any pairwise list of head words. We reject and re-sample sets with pairs above the threshold.

Design and procedure   Figure 2 illustrates the design. Each game contained 5 blocks, and each block had 6 trials. All trials in a game were based on the same context. In each context, 5 tangrams were assigned to the *repeated* condition, each appearing exactly once in each block. Among the 5 repeated tangrams, we ensured that 2 were drawn from the low-nameability condition and 3 from the high-nameability condition, or vice versa. The other 5 tangrams were assigned to the *control* condition and only appear in one of the blocks as the target. Thus, all experimental manipulations were within-dyad. The order of targets are randomized in each block, and participants alternated roles between blocks.

Participants were randomly assigned to pairs after providing consent and passing a tutorial and a quiz. The participants were randomly assigned to the speaker or the listener roles. When the game started, both players saw 10 tangrams and a chat box. The order of the tangrams was different from the speaker's and the listener's views, so that the speaker could not rely on the position when describing the target. The speaker was asked to send only one message to describe the highlighted target tangram from the context in 45 seconds. The listener needed to select a tangram based on the description. An additional 15 seconds were given to the listener to make the selection if they had not already. Both participants received feedback after each trial indicating if the listener had responded correctly. At the end of the experiment, the participants were given a demographic survey about their age, gender, language background, game experience, feedback for the study, etc. The experiment was built with Empirica (Al-maatouq et al., 2020), a platform for building and conducting synchronous and interactive online experiments with human participants. We hosted the games on Meteor Cloud and stored game data in MongoDB.

Results

We tested three key hypotheses about performance over time. First, we expected that high-nameability targets would be "easier" to communicate about than low-nameability targets across the board. Second, we expected that performance would improve overall for repeated targets to a greater degree than the non-repeated controls interspersed throughout the trial sequence. Third, and of greatest theoretical interest, we ask whether there is a three-way interaction: the performance improvements that accrue over successive interactions for repeated targets may more readily transfer to improved performance on low-nameability controls than high-nameability controls.

We evaluated these predictions using a series of mixed-effects regression models for three complementary metrics of communication performance: referential accuracy (whether or not the listener successfully selected the target), verbosity
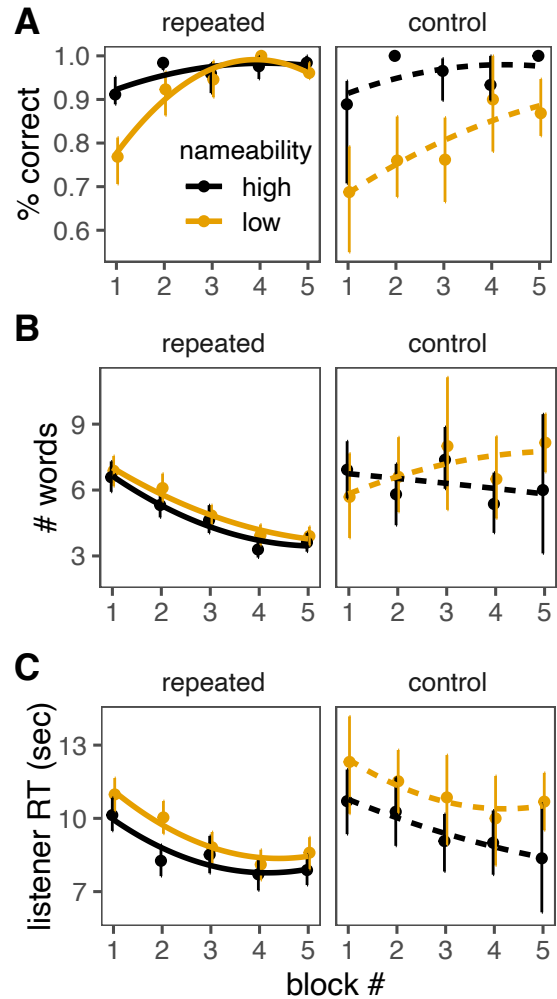
Figure 3: **Experiment 1 results**. (A) accuracy, (B) description length, and (C) time elapsed between speaker message and listener response (in seconds) for low- and high-nameability tangrams in the repeated (left) or control (right) condition. Error bars are bootstrapped 95% CIs.

(the number of words in the speaker's description), and speed (the time taken for the listener to make a selection after receiving the message). For each metric, we construct a regression model including fixed effects of condition (repeated vs. control) and nameability class (high vs. low), as well as an effect of block number (integers 1 to 5, centered) and all their interactions. For accuracy (coded as a binary variable: correct vs. incorrect), we use a logistic linking function. Because all manipulations were within-dyad, we included the maximal random effect structure at the dyad-level.

Accuracy   Our raw accuracy metric is shown in Figure 3A. First, we observed a significant main effect of nameability: all else equal, listeners were more likely to make errors for low-nameability targets than high-nameability ones, $b = 0.87, z =$
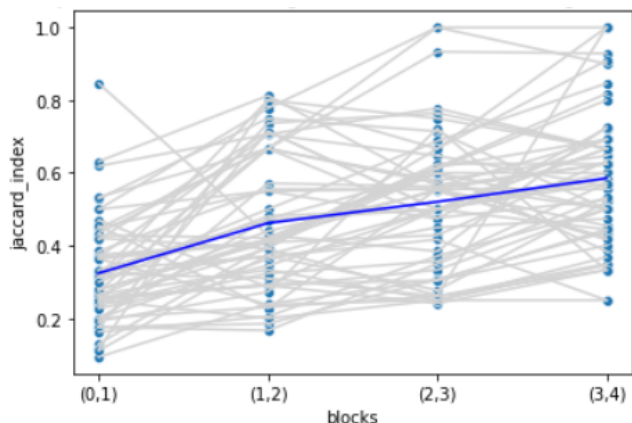
Figure 4: **Experiment 1 analysis**. To evaluate the stability of referring expression content over the course of the game, we computed the Jaccard index between the set of all words used in each pair of successive blocks. We found a gradual increase in the metric, indicating increasing block-to-block similarity in word sets.

$-3.4, p < 0.001$. While success rates improved overall over the course of the task, $b = 33.3, z = 4.1, p < 0.001$, the most complex model supported by our data only included a significant interaction between condition and nameability, $b = -0.32, z = -2.06, p = 0.039$, likely reflecting ceiling effects for high-nameability targets, and no significant interaction was found with block number.

Description length We considered the efficiency with which speakers were able to communicate, measured as the number of words in their descriptions (Figure 3B). We first found an overall main effect of block, $b = -18.6, t(105) = -3.7, p < 0.001$, consistent with classic observations (Clark & Wilkes-Gibbs, 1986; Krauss & Weinheimer, 1964), as well as a significant main effect of nameability, $b = 0.38, t(47) = 2.8, p = 0.008$, where high-nameability targets received shorter descriptions. We also found a significant interaction between condition and block: holding nameability constant, the decrease in utterance length for repeated tangrams was significantly greater than for control tangrams, $b = 25.2, t(689) = 6.5, p < 0.001$. Finally, we found a significant three-way interaction, $b = 8.4, t(1066) = 2.1, p = 0.039$, consistent with the hypothesis that dyads are able to converge on shorter descriptions for repeated tangrams regardless of their nameability, but these improvements only generalize to high-nameability control tangrams. If anything, when speakers must refer to a low-nameability control tangram for the first time late in the game, they produce slightly longer expressions than they did earlier in the game.

Listener response time Our third performance metric is the time required for listeners to respond after receiving a description (Figure 3C), as listeners may be expected to pause

longer when uncertain about their response. To ensure that listener response times are not driven by the time window left by speakers, we capped response times at the maximum value of 15 seconds given to listeners on trials when the speaker ran out the clock. We found a significant main effect of nameability, where listeners responded more quickly overall for high-nameability targets, $b = 0.63, t(139) = 4.7, p < 0.001$. As with accuracy, we did not find any significant interactions with block number, but did find an interaction between condition and nameability, $b = 0.28, t(1372) = 2.3, p = 0.02$. Although listeners were always a bit slower to respond for control tangrams than repeated tangrams, the gap was significantly bigger for low-nameability ones.

Stability of descriptions So far we have observed that dyads are able to communicate more efficiently and accurately over the course of a game, as a function of nameability. However, these coarse metrics are agnostic to the actual linguistic content of descriptions, the vocabulary used by the speaker in order to describe targets. To examine whether interlocutors are truly converging on shared conventions (despite swapping roles each block), we calculated a measure called the Jaccard index, $J(W_i, W_{i+1})$, between the set of words W_i used on successive blocks for tangrams in the *repeated* condition. The Jaccard index is a measure of set similarity defined as the size of the intersection over the size of the union, e.g.

$$J(W_1, W_2) = |W_1 \cap W_2| / |W_1 \cup W_2|$$

To ensure effects were not driven by spurious changes in function words or pluralization, we lemmatized all descriptions and excluded stop words prior to computing sets. The results of this analysis are shown in We found a steady increase in similarity across successive blocks Figure 4, indicating that participants increasingly reused lexical choices from their partner in the previous block as shared conventions stabilized.[1]

Experiment 2: Generalizing to unseen targets

Our findings from Experiment 1 not only emphasize the added difficulty of communicating about low-nameability objects overall, but also hint at the additional difficulty of *generalizing* newly acquired conventions to low-nameability objects. This effect was particularly strong for description length, where speakers were only willing to extend reduced descriptions to high-nameability control objects. One explanation for the differential effect of nameability on control trials is that speakers became increasingly confident that their partner would share the same meaning for unseen tangrams only when they were expected to have high consensus *a priori*

---

[1] We found a similar effect for a directed variant of the Jaccard index: $J_{dir}(W_1, W_2) = |W_1 \cap W_2| / |W_2|$ normalizing by the size of the later description rather than the union. This supplemental analysis suggests that the observed changes in set similarity were not driven by non-stationarity in set size over time.

Experiment 1

New Testing Context

High nameability ——— Low nameability

Testing

a candle in a holder
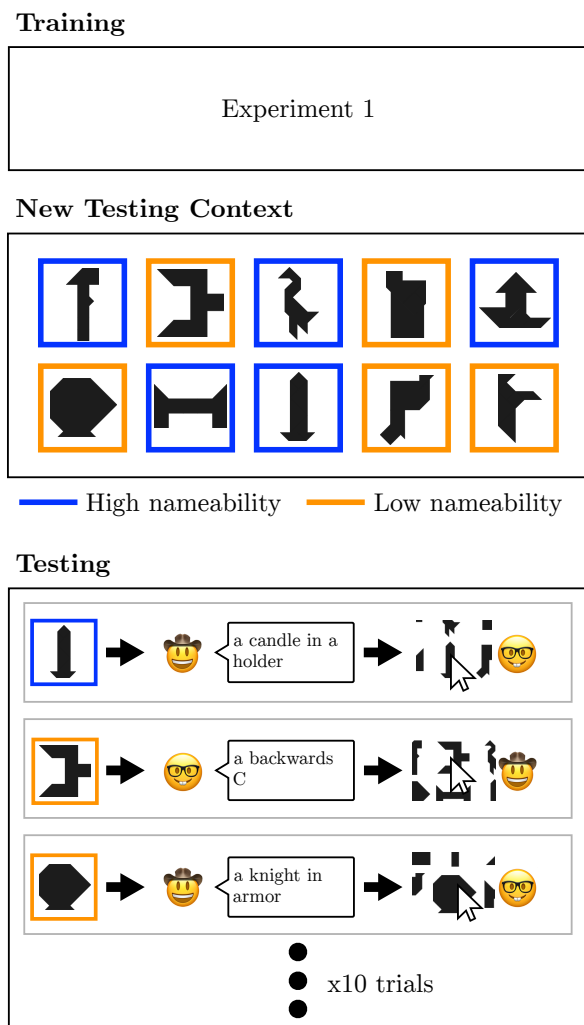
a backwards C

a knight in armor

x10 trials

Figure 5: **Experiment 2 design.** After participants completed the same procedure as Experiment 1 as a *training phase*, they proceeded to a new *test phase* where they played a reference game with an entirely new context of 10 tangrams.

and existing conventions were likely to be effective (Murthy et al., 2022). However, it is also possible that these control tangrams were simply more *familiar* as they had appeared in context alongside the repeated tangrams prior to appearing as the target. In Experiment 2, we introduce a stronger test of this hypothesis by measuring how speakers generalize to new contexts where all targets are entirely novel.

Methods

Participants We recruited 60 pairs of participants, 8 of whom were excluded based on the same criteria used in Experiment 1. Games lasted an average of 27 minutes and participants were paid $5.50 (approximately $11 per hour) with a performance bonus up to $1.20.

Stimuli, design and procedure We used the same sets of high- and low-nameability stimuli from Experiment 1, and
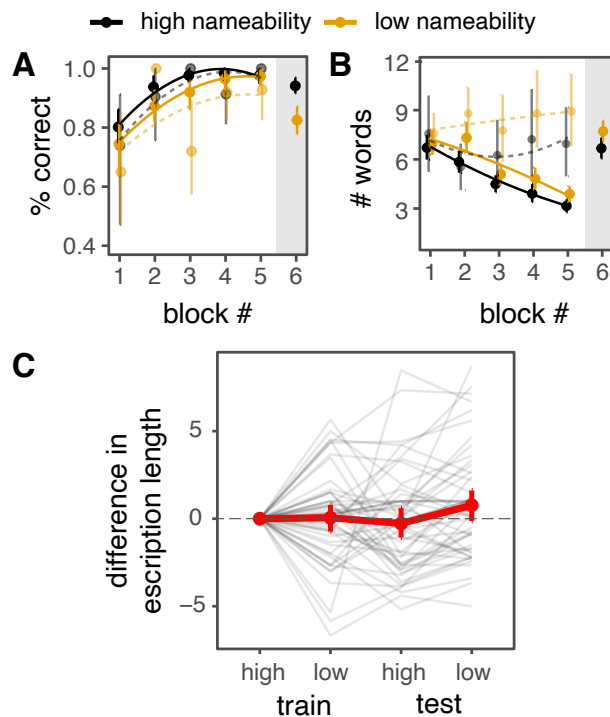

Figure 6: **Experiment 2 results**. (A) Accuracy and (B) description length for the train phase (blocks 1-5) and test phase (block 6). Transparent dashed lines are control conditions. (C) Within-game differences in description length relative to the first block of train for the first train and test blocks, for high- and low-nameability targets. Error bars are bootstrapped 95% CIs.

the procedure was a direct replication and extension. The first phase of the experiment (the *training* phase) was an exact replication of the within-dyad 2×2 design used in Experiment 1. The second phase (the *test* phase) was new. A 6th block was appended, containing 10 additional trials (Figure 5). Critically, these test trials used a completely non-overlapping context with 10 new targets presented in a randomized sequence. Each tangram in the new context was given a single trial. Speaker and listener roles were swapped between every trial.

Results

We evaluated the performance of a new context by examining two metrics: accuracy and verbosity (we omit the Experiment 2 response time analysis due to space constraints).

Accuracy In addition to replicating the nameability effects examined in Experiment 1 (see Figure 6A), the primary analysis of interest is a direct comparison against the initial block of the train phase (block 1) and the initial block of the test phase (block 6). In both cases, it is the speaker's first time referring to all targets in context. Thus, any differences in the test phase can be attributed to some generalizable learning

taking place over the training block. We ran a mixed-effects logistic regression model predicting accuracy only for these two blocks, including fixed effects of phase (train vs. test) and nameability (low vs. high), as well as random intercepts and slopes at the dyad-level. We found a significant interaction, $b = -0.31, z = 2.25, p = 0.024$, indicating that there was greater improvement from the beginning of train to test for high-nameability tangrams than low-nameability ones.

Description length   Average description lengths are shown for the training phase and test phase in Figure 6B. We again replicated the nameability effects found in Experiment 1, and focus here on the direct comparison between the first block of training and the first block fo test. We ran a mixed-effects linear regression model predicting description length, including fixed effects of phase (train vs. test) and nameability (low vs. high) with random intercepts and slopes at the dyad-level. In addition to a significant main effect of nameability at the beginning of both train and test, $b = 0.32, t(112) = -2.181, p = 0.03$, we found a marginal interaction, $b = 0.25, t(629) = 1.8, p = 0.069$. In a Bayesian mixed-effects model with full random effects, we obtained a 95% credible interval of $[-0.05, 0.54]$. In other words, there was a small but meaningful gap in description length between high and low-nameability tangrams in the test phase, while no such gap was observed in the first block (see Figure 6C for a finer-grained visualization controlling for individual differences in overall verbosity). Put together, these effects suggest that participants were better able to anticipate in the test phase which tangrams would be harder and adjust their description length accordingly.

Discussion

The study conducted in this paper investigated the effect of nameability on the process of reference generalization using tangrams from the KILOGRAM dataset, a large-scale high-diversity resource for abstract stimuli. We conducted two experiments that track the development of conventions during a repeated reference game, and assessed the impact of these conventions on a reference game with a new context. We observed that conventions, expressed through more efficient communication and stable vocabulary, are formed through repeating games, re-affirming observations from past work (Clark & Wilkes-Gibbs, 1986; Hawkins et al., 2020) using the larger-scale KILOGRAM data (Ji et al., 2022). Our key contributions concern the transfer of conventions to novel stimuli (i.e. control images), and the influence of nameability on this process. Our analyses showed that conventions formed for repeating stimuli transfer to both control stimuli that do not repeat, and to completely new context with previously unseen tangrams. This form of generalization has been generally understudied in prior work. Moreover, this process is modulated by the nameability of the target objects. Objects with higher nameability lent themselves to transfer of acquired conventions significantly more than objects with lower nameability.

Our studies raise several important directions for future work. Most importantly, we leave open the question of what, exactly, is transferred. One possibility is that pairs of participants aligned their semantic spaces in a way that generalized beyond the specific visual stimuli observed in the repeated condition. This conceptual alignment then allowed them to produce and comprehend references in new contexts more efficiently and more effectively. The differences we observed across nameability classes suggest that this alignment is not perfect, and generally noisier for objects that are harder to describe. Indeed, the conceptual alignment hypothesis predicts that generalization would fall off as a function of perceptual or conceptual distance. An alternative hypothesis is that participants developed a shared model of which stimuli are easy or hard to describe (i.e. have high or low nameability). The speakers could have then utilized this information and the mutual knowledge that it is shared with their partners to be more efficient referring to objects in the new context. In a nutshell, the two hypotheses contrast the direct tuning of semantic knowledge against the acquisition of a higher-level task model. Identifying which, if either, of these hypotheses is correct requires further studies, likely with more constrained structure. For example, rather than using behavior during the first block as a baseline, which also impacts the actual conventions formed, future studies could prefix the repeating game with an asocial naming block.

Another important direction for future work concerns how the statistics of the object distribution in each of the contexts (either for the train or test phase), may influence the transfer of conventions. For example, could contexts with a larger ratio of high nameability objects facilitate faster convention formation? Could contexts with more low nameability objects provide a more challenging environment and result in more enduring conventions? Our studies have also not explored the impact of the semantics of the objects on participants behavior. For example, if we were to create contexts that include more semantically related objects, would that lead to faster or slower convention formation? Would it result in more robust transfer to new contexts? Many of these questions could be readily addressed with carefully designed contexts making use of the diversity and scale of KILOGRAM.

It is highly unlikely for the exact same concepts and situations to repeat in realistic situations. Thus, generalizing partner-specific conventions to new concepts is an crucial part streamlining interaction with repeating partners. Understanding and modeling such generalization behavior is paramount to gain insight into the development of language as a system for coordination and efficient communication between individuals.

Acknowledgments

References

Almaatouq, A., Becker, J. P., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2020). Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, *53*, 2158–2171.

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

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482.

Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as contexts in conversation. In *Psychology of learning and motivation* (pp. 59–99).

Clark, H. H. (1996). *Using language*. Cambridge university press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*(2), 233–263.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4), 591.

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics uncertainty in language and thought. *The Handbook of Contemporary Semantic Theory*, 655–686.

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, *44*(6), e12845.

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2022). From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo.

Hupet, M., Seron, X., & Chantraine, Y. (1991). The effects of the codability and discriminability of the referents on the collaborative referring procedure. *British Journal of Psychology*, *82*(4), 449–462.

Ibarra, A., & Tanenhaus, M. K. (2016). The flexibility of conceptual pacts: Referring expressions dynamically shift to accommodate new conceptualizations. *Frontiers in Psychology*, *7*, 561.

Ji, A., Kojima, N., Rush, N., Suhr, A., Vong, W. K., Hawkins, R., & Artzi, Y. (2022). Abstract visual reasoning with tangram shapes. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 582–601.

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*.

Lewis, D. (1970). General semantics. *Synthese*, *22*(1-2), 18–67.

Lupyan, G., Uchiyama, R., Thompson, B., & Casasanto, D. (2023). Hidden differences in phenomenal experience. *Cognitive Science*, *47*(1), e13239.

Martı, L., Wu, S., Piantadosi, S. T., & Kidd, C. (2023). Latent diversity in human concepts. *Open Mind*, *7*, 79–92.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213.

Murthy, S. K., Griffiths, T. L., & Hawkins, R. D. (2022). Shades of confusion: Lexical uncertainty modulates ad hoc coordination in an interactive communication task. *Cognition*, *225*, 105152.

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93–104.

Potts, C., Lassiter, D., Levy, R., & Frank, M. C. (2015). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, *33*(4), 755–802.

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in cognitive sciences*.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Stolk, A., Verhagen, L., & Toni, I. (2016). Conceptual alignment: How brains achieve mutual understanding. *Trends in Cognitive Sciences*, *20*(3), 180–191.

Wang, X., & Bi, Y. (2021). Idiosyncratic tower of babel: Individual differences in word-meaning representation increase as word abstractness increases. *Psychological Science*, *32*(10), 1617–1635.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*(2), 183–194.

Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition*, *196*, 104135.