ORIGINAL ARTICLE

Characterizing the dynamics of learning in repeated reference games

Robert Hawkins[1] | **Michael Frank**[1] | **Noah Goodman**[1,2]

[1]Department of Psychology, Stanford University

[2]Department of Computer Science, Stanford University

Correspondence
Robert Hawkins, Department of Psychology, Princeton University, Princeton, NJ 08540
Email: robertdh@princeton.edu

Funding information

When talking with novel partners about novel referents, speakers must continually adapt to coordinate on meaning. Here we draw upon recent advances in natural language processing to provide a finer-grained characterization of the dynamics of such adaptation. We release an open corpus (>15,000 utterances) of extended dyadic interactions in a classic repeated reference game task where pairs of partners had to coordinate on how to talk about initially difficult-to-describe tangram stimuli. We find that different pairs coordinate on a wide range of idiosyncratic solutions to the problem of reference, but they do so in a highly systematic manner. Words that are more discriminative in the initial context (i.e. that were used for one target more than others) are more likely to persist through the final round. Structurally, we found that utterances reduce across pairs following similar trajectories: entire modifying syntactic units drop out as conventions are established, leaving only contentful open-class parts of speech. These findings provide higher resolution into the quantitative dynamics of *ad hoc* convention formation: based on a shared history of usage, words systematically acquire new meaning with a partner to support more efficient communication.

KEYWORDS
social coordination, conventions, semantics, syntax, natural language processing, semantic embeddings

1 | INTRODUCTION

Human language use is remarkably flexible. We are able to coax new meanings out of existing words, or even coin new ones, to handle the diverse challenges encountered in everyday communication (Clark, 1983; Davidson, 1986). This flexibility is partially explained by one-shot pragmatic reasoning, which allows listeners to use context to infer an intended meaning even in cases of ambiguous or non-literal usage (Lascarides and Copestake, 1998; Glucksberg and McGlone, 2001; Piantadosi et al., 2012; Goodman and Frank, 2016). However, a rich theoretical thread in the literature has suggested that *learning* mechanisms may also play an important role, allowing speakers and listeners

to dynamically adapt their representations of meaning over the course of an interaction (Brennan and Clark, 1996; Pickering and Garrod, 2004; Delaney-Busch et al., 2019).

Two functional considerations motivate the need for continued learning in communication, even among adults. First, just as there is substantial phonetic variability across speakers with different accents (Kleinschmidt, 2019), lexical items may vary in meaning from speaker to speaker. This variability is clear for cases like slang, technical lingo, nicknames, or colloquialisms (e.g. Clark, 1998), but may extend even to more ordinary nouns and adjectives. It may be difficult to know at the Second, because we live in a changing environment, we often experience novel entities, events, thoughts, and feelings we want to talk about but do not already have (literal) words to express. Both of these obstacles can be overcome using feedback from one's partner to dynamically re-calibrate expectations about meaning.

The *repeated reference game* task has provided a natural and productive paradigm for eliciting behavior under such conditions. In this task, pairs of participants are presented with arrays of novel, conceptually ambiguous objects. On each trial, one player – the speaker – is privately shown a *target* object and must produce a referring expression allowing their partner to correctly select that object from the context. The speaker is then given feedback at the end of each trial about which object the listener selected, and the listener is given feedback about the true target object. Critically, each object appears as the target multiple times in the trial sequence, allowing the experimenter to examine how referring expressions change as the speaker and listener accumulate shared experience. To the extent that the speaker and listener converge on an accurate system of stable referring expressions, and these referring expressions differ from the ones that were initially produced, the data indicate that *conventions* or *pacts* have formed through some learning process.

Indeed, a key early result first reported by Krauss and Weinheimer (1964) is that descriptions are dramatically shortened across repetitions: an initial description like "the one that looks like an upside-down martini glass in a wire stand" may gradually converge to "martini" by the end. Subsequent work has established a number of signature properties of convention formation through careful experimental manipulation. For example, reduction is contingent on listener feedback (Krauss and Weinheimer, 1966; Krauss et al., 1977; Hupet and Chantraine, 1992), and is therefore not easily explained as a mere practice or repetition effect; the resulting conventions are *partner-specific* in the sense that they do not transfer if a novel listener is introduced (Wilkes-Gibbs and Clark, 1992; Metzing and Brennan, 2003; Brennan and Hanna, 2009); and they are *sticky* in the sense that they persist through precedent even after the context changes (Brennan and Clark, 1996).

A variety of models, at varying levels of formal precision, have been proposed to explain these qualitative effects. However, further model development depends critically upon a finer-grained characterization of the *quantitative* signatures of adaptation. Certain fundamental descriptive questions remain unanswered, and important theoretical constructs remain poorly operationalized. For example, while it has been widely observed that utterances reduce in length as common ground is accumulated, a precise characterization *what* gets reduced, and *how*, has remained elusive. What determines whether a particular word is dropped or preserved? Are words dropped randomly or clustered together in phrases? Similarly, while constructs like arbitrariness or stability have loomed over the theoretical analysis of conventions (e.g. Lewis, 1969), it has been unclear how exactly to measure the extent to which these properties hold in a particular task and how they may evolve over the course of interaction. Without addressing these gaps in measurement, it is difficult to set finer-grained criteria to distinguish among different models.

In this paper, rather than arguing for a particular model, we release a large corpus of repeated reference games and conduct a variety of computational analyses to provide a foundation for future model development. The computational techniques necessary to analyze such rich natural language data were limited at the time of prior work, but have become newly tractable given developments in natural language processing (NLP). Our analyses divide into two broad categories roughly corresponding the dynamics of *content* and *structure* of referring expressions across

interaction. To examine content, we extracted word embeddings (e.g. GloVe vectors) for each message to calculate the similarity of messages within and across pairs. We found that while different pairs coordinate on a wide range of idiosyncratic solutions to the problem of reference, they do so in an increasingly stable and path-dependent manner. Further, words that are more discriminative in the initial context (i.e. that were used for one target more than others) are more likely to persist through the final round. To examine structure, we extracted parts of speech and syntax trees from the text to understand what was reducing and how. We found that pairs systematically drop entire modifying phrases at each repetition, leaving only open-class parts of speech (e.g. an adjective and noun) by the final round. These findings provide a new window into the quantitative dynamics of learning in communication and raise new questions about how conventions form.

2 | METHODS: REPEATED REFERENCE EXPERIMENT

To collect a large corpus of natural dialogue that allows us to measure how pairs coordinate on meaning over time, we faced two primary decisions. First, to observe the formative period of linguistic conventions, we required novel, ambiguous stimuli for which participants didn't already have strong initial conventions. Second, to observe the *dynamics* of conventions over time, we needed the same coordination problem to be repeated over time, such that earlier outcomes are relevant for later decisions. These criteria are satisfied by a *repeated reference game* design in which participants refer to the same objects across multiple rounds as they build up a shared history of interaction, or common ground, with their partner.

We developed two variants of the game: a relatively unconstrained *free-matching* version that more closely replicates the classic in-lab design, and a more tightly controlled *cued* version that allows for higher resolution analyses of how references to individual tangrams changed over time (see Fig. 1). The *free-matching* version was an exploratory sample, but we pre-registered our full pre-processing and analysis pipeline for the *cued* version[1]. While we have released the corpus from the free-matching version[2], we only use the *cued* version throughout the paper as our confirmatory sample.

2.1 | Participants

A total of 480 participants (218 in the *free-matching* version and 262 in the *cued* version) were recruited from Amazon's Mechanical Turk and paired into dyads to play a real-time communication game using the framework in Hawkins (2015).

2.2 | Exclusion criteria

After excluding games that terminated before the completion of the experiment due to server error or network disconnection (40 in *free matching* and 33 in *cued*), as well as games where participants reported a native language different from English (2 in *free matching* and 3 in *cued*), we implemented an additional exclusion criterion based on accuracy. We used a 66/66 rule, excluding pairs that got fewer than 66% of the tangrams correct ($\geq$ 8 of 12) on more than 66% of blocks ($\geq$ 4 of 6). While the most pairs were near ceiling accuracy by the final round, this rule excluded 11 in *free matching* and 8 in *cued* who appeared to be guessing or rushing to completion. After all exclusions, we were left with

[1] https://osf.io/2zwmx

[2] https://cocolab.stanford.edu/datasets/tangrams.html

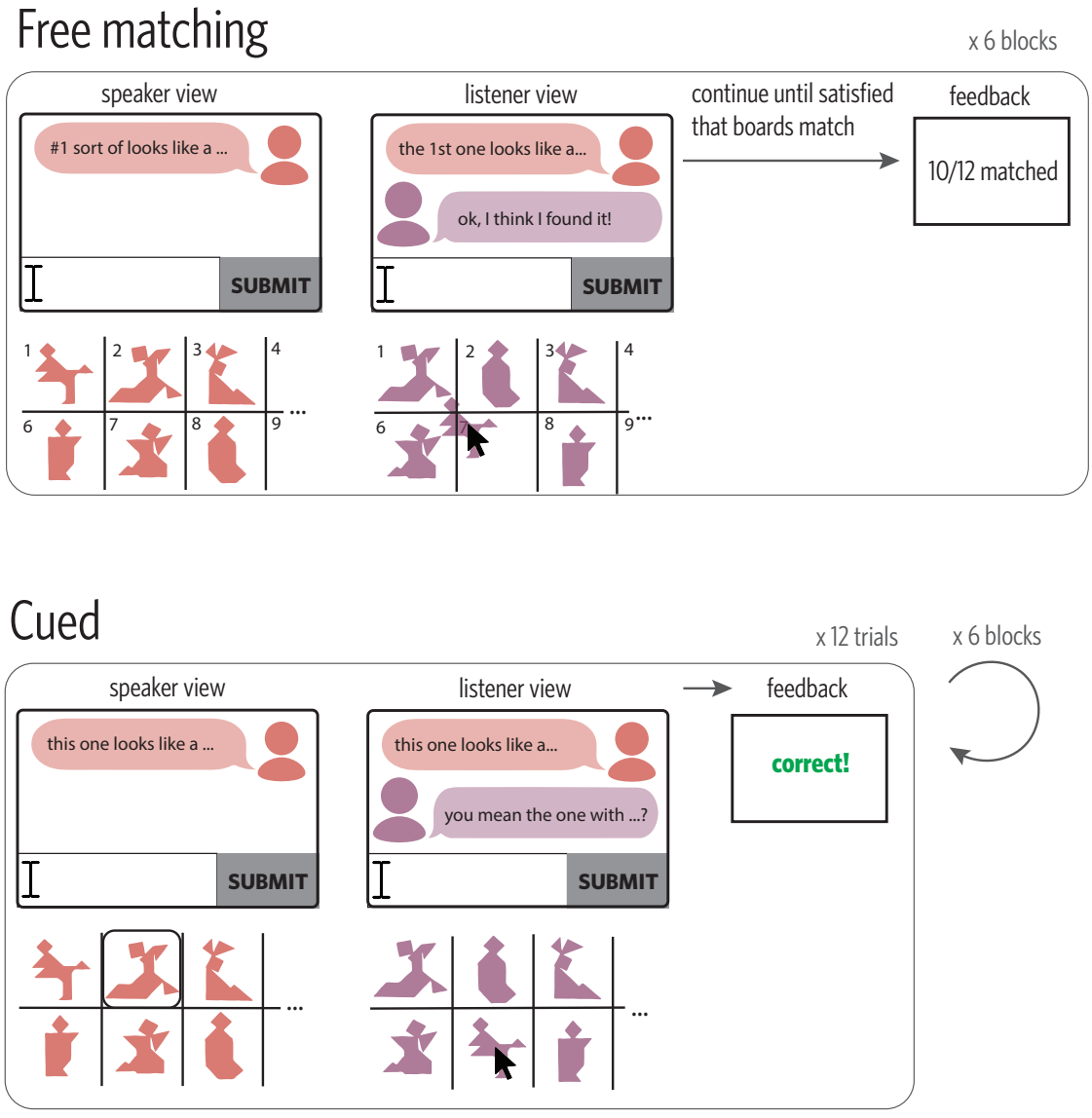

FIGURE 1 'Free matching' and 'cued' variants of the tangrams task.

a *free matching* corpus containing a total of 8,639 messages over 56 complete games and a *cued* corpus containing 9,164 messages over 83 games.

2.3 | Stimuli & Procedure

On every trial, participants were shown a 6×2 grid containing twelve tangram shapes, reproduced from Clark and Wilkes-Gibbs (1986). After passing a short quiz about task instructions, participants were randomly assigned the role of either 'director' or 'matcher' and automatically paired into virtual rooms containing a chat box and the grid of stimuli. Both participants could freely use the chat box to communicate at any time.

In the *free-matching* version, our procedure closely followed Clark and Wilkes-Gibbs (1986). The director and matcher began each round with scrambled boards. The director's tangrams were fixed in place, but the matcher's could be clicked and dragged into new positions. The players was instructed to communicate through the chat box such that the matcher could rearrange their shapes to match the order of the director's board. When the players were satisfied that their boards matched, the matcher clicked a 'submit' button that gave players batched feedback on their score (out of 12) and scrambled the tangrams for the next round. After six rounds, players were redirected to a short exit survey. Cells were labeled with fixed numbers from one to twelve in order to help participants easily refer to locations in the grid (see Fig. 1).

While this replicated design allows for highly naturalistic interaction, it poses several problems for text-based analyses. First, utterances must contain not only descriptions of the tangrams but also information about the intended location (e.g. '*number 10* is the …'). Additionally, because there were no constraints on the sequence, participants can revisit tangrams out of order or mention multiple tangrams in a single message, making it difficult to isolate exactly which utterances referred to which tangrams without extensive hand-annotation. Finally, the design of the 'submit' button made it easy for players to occasionally advance to the next round without referring to all 12 tangrams.

For the *cued* version, then, we designed a more straightforwardly sequential variation on the task where speakers are privately cued to refer to targets one-by-one and feedback is given on each round (see Fig. 1); this allows us to straightforwardly conduct analyses at the tangram-by-tangram level. On each trial, one of the twelve tangrams was privately highlighted for the director as the *target*. Instead of clicking and dragging into place, matchers simply clicked the one they believed was the target. They were not allowed to click until after a message was sent by the speaker. We constructed a sequence of six blocks of twelve trials (for a total of 72 trials), where each tangram appeared once per block. Because targets were cued one at a time, numbers labeling each square in the grid were irrelevant and we removed them. The context of tangrams was scrambled on every trial, and participants were given full, immediate feedback: the director saw which tangram their partner clicked, and the matcher saw the intended tangram.

2.4 | Data pre-processing

We used a three step pre-processing pipeline to prepare our corpus for subsequent analyses. Unless otherwise noted, we used the open-source Python package `spaCy` to implement all NLP tasks.

1. **Spell-checking and regularization**: We conservatively extracted all tokens that did not exist in the vocabulary of the smallest available (~ 50,000 word) `spaCy` model and passed them through the SymSpell spell-checker [3]. These suggested corrections were then sequentially presented to the first author and either accepted or overridden at their judgement. This process constructed a reproducible spell-correction dictionary we applied to our dataset.

[3] `https://github.com/wolfgarbe/SymSpell`

2. **Cleaning unrelated discourse**: Because we allowed our participants to interact in real-time through the chat box, many pairs produced text unrelated to the task of referring to the current target (e.g. greeting one another, asking personal questions, commenting on the length of the task or the results of previous rounds). We wanted to ensure that our structural results were not confounded by patterns in this kind of discourse across the task, and that the semantic content we observe on a particular trial is in fact being used to refer to the current target rather than task-irrelevant topics or, as we found in some cases, referring to other tangrams while debriefing previous errors. We therefore applied a manual pass applying a rubric that any text not directly referring to the current target is removed. For example, utterances like "this is the one we got wrong last time" were kept in because they were referring to a property of the current tangram, but utterances like "good job" and "they'll go quicker if you remember what I say!" are not. This process also created a reproducible JSON.
3. **Collapsing multiple messages within a round**: Finally, some speakers used our chat box like an texting interface, hitting the enter key between every micro-phrase of text. This made it difficult to interpret the output of syntactic parses. We therefore collapsed repeated messages by a participant within a round into a single message by inserting commas between successive messages. We chose to use commas because it tends to maintain grammaticality and does not inflate word counts.

3 | RESULTS: CHARACTERIZING THE DYNAMICS OF CONTENT

Based on recent developments in theory, we pre-registered about how speakers change the content of their referring expressions over time. *First*, if participants are influenced on each trial by pragmatic pressures to be informative, the labels that conventionalize should not be a random draw from the initial description. Instead, we predict that more *distinctive* words in initially successful labels (e.g. words used exclusively to describe one tangram) will be more likely to remain in later descriptions. *Second*, due to sources of variability in the population of speakers, we predict that the referring expressions used by different pairs will increasingly diverge to different, idiosyncratic labels. In other words, different pairs will find different but equally successful equilibria in the space of possible linguistic conventions. *Third*, as speakers learn and gradually strengthen their expectations about how their partner will interpret their referring expressions, the labels used within each pair for each tangram with stabilize. In other words, once there is evidence that a particular label is successfully understood, there is little reason to deviate from it. Because these analyses depend on tangram-level resolution, we only examine the "cued" dataset in this section.

3.1 | Initially distinctive words are more likely to conventionalize

We begin by investigating *which* content is dropped and which is preserved. Two computational principles guide our exploration of this question. First, if speakers are attempting to be informative in a particular context of other tangrams then the Gricean maxim of quality suggests that a good referring expression is one that applies more strongly to the target than to the distractors. Properties that are shared in common across multiple objects are poor candidates for conventions that must distinguish among them. Second, the principles of cross-situational learning suggest that these informativity considerations will be strengthened over time. The exclusive usage of a word with one tangram and no others should reinforce the specificity of that meaning in the local discourse context, even if the listener may be *a priori* willing to extend it to other targets. Conversely, if a particular word has been successfully used with several different referents, its specificity may be weakened in the local context. Putting these principles together, we hypothesized that more *initially distinctive* words would be more likely to conventionalize.

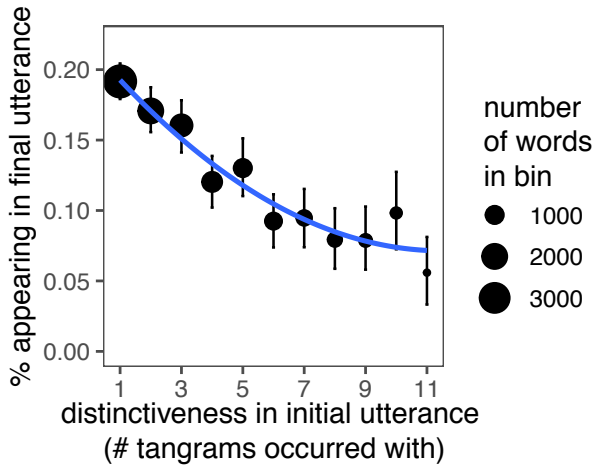

FIGURE 2 More distinctive words are more likely to conventionalize. Points represent estimates of the mean probability of conventionalizing across all words with a given distinctiveness value. Size of points represent the number of words at that value. Curve shows regression fit; error bars are bootstrapped 95% CIs.

For each pair of participants, we quantified the distinctiveness of a word w as n_w: the number of tangrams that it was used to describe on the first repetition. A word that is only used in the description of a single tangram (e.g. a descriptive noun like "rabbit") would be very distinctive, while a word used with all 12 tangrams (e.g. an article like "the") would be not distinctive at all. While this formulation is the most transparent to state in words, it is equivalent (up to a constant) to two popular and theoretically motivated measures of distinctiveness used in natural language processing (Salton and Buckley, 1988). The first is *term frequency-inverse document frequency* (tf-idf, Sparck Jones, 1972), which multiplies the term frequency $tf(w, d)$ of a word w in a document d by a "global" term $\log(N/n_w)$ where N is the total number of documents and n_w is the number of documents containing w. In our case, the "documents" are just the referring expressions used for a distinct tangram on the first round, so $N = 12$ and we can take $tf(w, d)$ to be a boolean for simplicity: 1 if the word occurs, 0 if it does not. We can thus retrieve our simpler measure by exponentiating, dividing by 12, and taking the inverse. The second is *positive point-wise mutual information* (PPMI). Point-wise mutual information compares the joint probability of a word occurring with a particular tangram to the probability of the two occurring independently:

$$PMI_{\text{word,tangram}} = \log \frac{P(\text{word, tangram})}{P(\text{word})P(\text{tangram})}$$

Positive point-wise mutual information is given by $\min(0, \text{PMI})$, restricting the lower bound to 0. It can be shown for our case that *tf-idf* is the maximum likelihood estimator for PPMI: the numerator reduces to a boolean when we only have one observation per tangram (Robertson, 2004).

Given this simple but principled measure of word distinctiveness at the speaker-by-speaker level – the number of tangrams it was initially used with – we were interested in the extent to which it accounts for conventionalization, the probability that a word in the speaker's initial description is preserved until the end of the game. More than half of the words used to refer to a tangram on the final round (57%) appeared in the initial utterance. [4] We thus

[4] The 43% of final round words that did not exactly match were often synonyms or otherwise semantically related to words used on the first round, e.g. "foot"

restricted our attention to this subset of words, coding them with a 1 if they later appeared in the final round and 0 if they did not. We then ran a mixed-effects logistic regression including a fixed effect of initial distinctiveness (using the transformed *tf-idf* for stability) and maximal random effect structure with intercepts and slopes for each tangram and pair of participants. We found a significant positive effect of distinctiveness: words that were used with a larger number of tangrams on the first round were less likely to conventionalize, $b = -0.23, z = -6.1$ (see Fig. 2). Similar results are found using the derived measure of *tf-idf*.

Finally, we conducted a non-parametric permutation test. For each speaker and tangram, we *randomly sampled* a word from the initial utterance and computed the mean probability of this word also being used on the final round. Repeating this procedure 1000 times yielded a null distribution ranging from 2.5% to 6.6%. However, if we instead sample from the words with *maximal distinctiveness*, we obtained a distribution ranging from 24% to 31%, which is non-overlapping with the null distribution. Thus, if we must make a bet on which words will become conventionalized, placing our bet on the most distinctive ones will yield much higher returns.

3.2 | Conventions diverge across pairs and stabilize within pairs

To jointly examine our other predictions about the dynamics of content, we introduce two different quantitative measures of similarity: one based on properties of the discrete word count distribution and the other based on distances computed between continuous vector embeddings of referring expressions. Because these analyses depend on the identity of word tokens, we applied a lemmatizer to further standardize the input. Lemmatization maps multiple morphological variants (e.g. 'played,' 'playing,' 'plays') to the same stem ('play'). We do not want an observed difference between two pairs to be driven simply by different forms of the same word.

Measuring convergence and divergence with discrete word distributions

We begin by examining the discrete *distribution of words* that each pair uses to refer to each tangram, excluding standard stop words. If a pair of participants converges on stable labels for a tangram, then this stability should manifest in a highly structured distribution over words throughout the game for that pair. If different speakers discover diverging conventions, this idiosyncracy should also manifest in differing word distributions. We formalize these intuitions by examining the information-theoretic measure of entropy:

$$H(W) = \sum_{w} P(w) \log P(w)$$

The entropy of the word distribution for a pair is maximized when all words are used equally often and declines as the distribution becomes more structured, i.e. when the probability mass is more concentrated on a subset of words.

To compare word distributions across games, we use a permutation test methodology. By scrambling referring expressions for each tangram across games and recomputing the entropy of the scrambled word distribution, we effectively disrupt any structure within each pair. There are two important inferences we can draw from this test. First, in a null scenario where different pairs did *not* diverge as predicted and instead every pair coordinated on roughly the same (optimal) convention for each tangram, this permutation operation would have no effect since it would be mixing together copies of the same distribution. Second, in another null scenario where pairs did not converge and instead varied wildly in the words they used from round to round, then permuting across games would also have no effect since it would simply mix together word distributions that already have high entropy. Hence, scrambling should *increase*

on the first round vs. "leg" on the last. In other cases, the labels used at the end were introduced after the first repetition, e.g. one pair only started using the conventionalized label "portrait" on repetition 3.

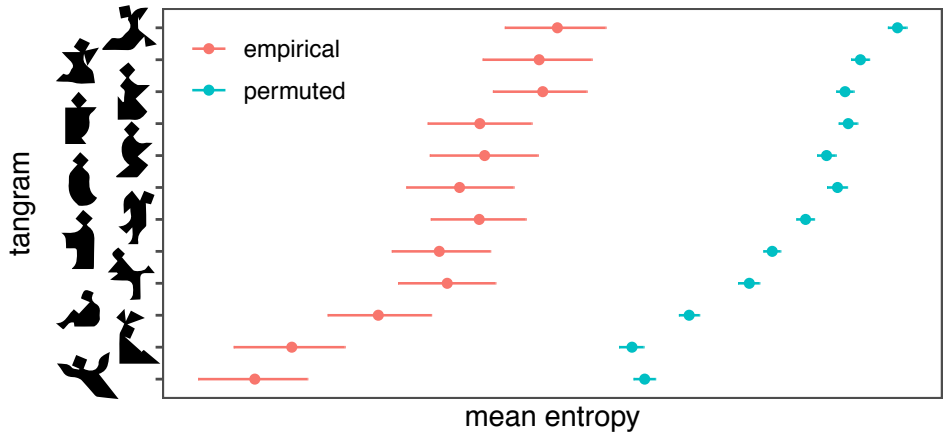

FIGURE 3 Permuting utterances across pairs increases entropy of word distribution, consistent with internal stability and multiple equilibria. Mean empirical entropy (red) and mean permuted entropy (blue) are shown for each tangram. Error bars are 95% CIs for bootstrapped empirical entropy and the permuted distribution, respectively.

the average game's entropy only in the case where both predictions hold: each game's idiosyncratic but concentrated distribution of words would be mixed together to form more heterogeneous and therefore high-entropy distributions.

Following this logic, we computed the average within-game entropy for 1000 different permutations of speaker utterances. We permuted utterances within rounds rather than across the entire data set to control for the fact that earlier rounds may generically differ from later rounds. Because we are permuting and measuring entropy at the tangram-level, this yields 12 permuted distributions (see Fig. 3). We found that the mean empirical entropy lay well outside the null distribution for all twelve tangrams, $p < .001$, consistent with our predictions of internal stability within pairs and multiple equilibria across pairs.

Measuring similarity using vector space embeddings

A more direct way to quantify convergence within and divergence across different pairs is to use a continuous similarity between vector space embeddings of utterances. Although the idea of using dense vector space representations of words to measure similarity is an old one (Osgood, 1952; Landauer and Dumais, 1997; Bengio et al., 2003), recent breakthroughs in machine learning have yielded rapid improvements in these representations (e.g. Mikolov et al., 2013; Pennington et al., 2014). To quantify the dynamics of semantic context in referring expressions across and within games, we extracted the 300-dimensional GloVe vector for each word. We then averaged these word vectors to obtain a single sentence vector for each referring expression[5]. To avoid artifacts from function words, we only included open-class content words (nouns, adjectives, verbs) in this average. We can then define a similarity metric between any pair of vectors $\langle u_i, u_j \rangle$. We find that our results are robust to several choices of metric, but for simplicity we will use cosine similarity

$$\cos\theta_{ij} = \frac{u_i \cdot u_j}{\|u_i\| \|u_j\|}$$

[5]Variations on such naive averaging methods are surprisingly strong baselines for sentence representations (Arora et al., 2017), performing better than supervised LSTM representations or unsupervised skip-thought vectors (Kiros et al., 2015)

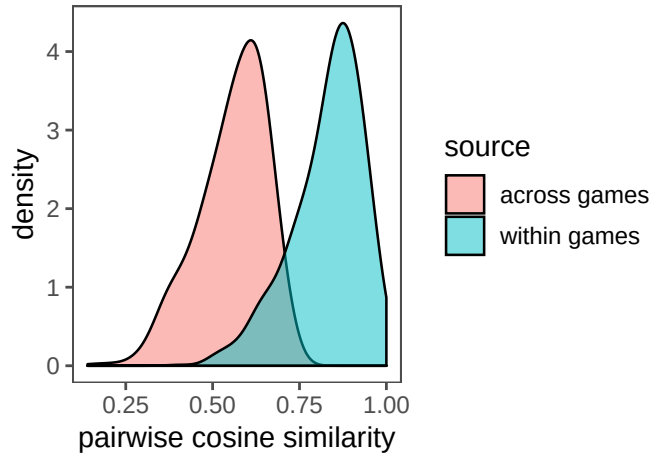

FIGURE 4 Distribution of similarities between different utterances within and across different games.

throughout the presentation below.

Utterances are more similar overall within games than between games

Before examining the dynamics of these vectors, we first test the basic prediction that utterances *within* a game are more similar overall than utterances *across* games, reflecting systematic variability in how different pairs solve the referential challenge posed by the reference game. For each tangram, we computed the pairwise similarity between all utterances *within* a game and also *across* games. The distributions of these values are shown in Fig. 4. We estimated the distance between these distributions using the standard normalized sensitivity $d' = \frac{\mu_A - \mu_W}{\sqrt{1/2(\sigma_A^2 + \sigma_W^2)}} = 2.71$. To compare this estimated difference against the null hypotheses that within- and across-game similarities are drawn from the same distribution, we conducted a permutation test by scrambling 'within' and 'across' labels for each similarity and re-computing d' 1000 times. We found that our observed value was extremely unlikely under this null distribution, 95% CI : $[-0.09, 0.09]$, $p < 0.001$.

In other words, utterances from a single pair tend to cluster together in semantic space while different pairs are more spread out in different parts of the space. This observation is consistent with our hypothesis that different pairs discover different conventions while a single pair tends to keep using a convention once established. Having established this separation between similarity distributions in aggregate, we proceed to ask more fine-grained questions about the *dynamics* through space: how do individual pairs evolve in their content over successive rounds? To more rigorously test our predictions about gradual divergence to multiple equilibria and convergence to internally stable conventions, we conducted three analyses directly on the semantic vectors.

Increasing dissimilarity from initial utterance

First, we hypothesized that there was cumulative change in the semantic content of a particular pair's utterances across repetitions. Concretely, we predicted that within a particular pair of participants, utterances on later repetitions would become increasingly dissimilar from the initial utterance. We tested this prediction in a mixed-effects regression model including (orthogonalized) linear and quadratic fixed effects of the 'lag' from the first repetition (i.e. 1 for the second repetition, 2 for the third repetition, etc) as well as maximal random effects for each tangram and pair

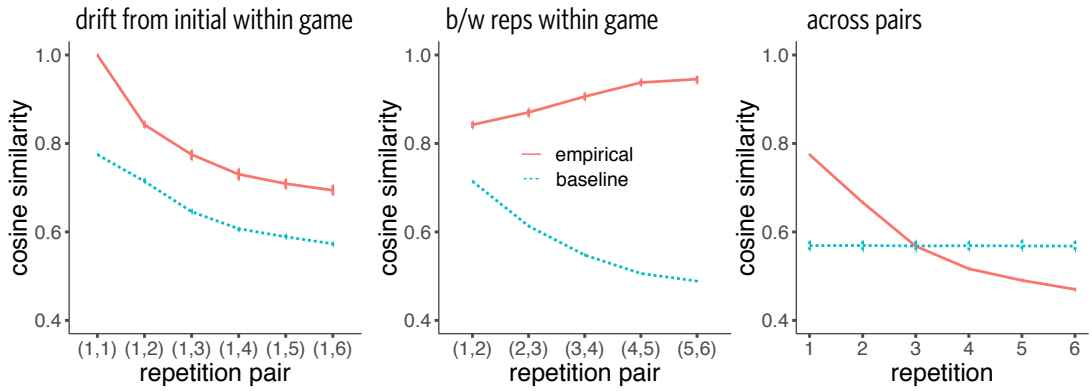

FIGURE 5 Utterances within a pair (A) become more dissimilar from initial utterance and (B) become more similar to successive utterances on later repetitions, but (C) utterances across pairs become steadily more dissimilar. Error bars are bootstrapped 95% CIs; dotted line represents permuted baseline.

of participants. We found a significant linear decrease in similarity to the initial round as the lag becomes larger, $b = -3.5$, $t = 12.2$, as well as a significant quadratic term, $b = 1.1$, $t = 5.3$, suggesting that this decrease in similarity slows down over time (see Fig. 5A).

However, since the entire distribution of utterances may have drifted to a different region of the semantic space for generic reasons (e.g., because they were shorter overall), we compared the estimated drift *within* pairs of participants to a permuted baseline. For each target tangram, we scrambled utterances across different pairs of participants and re-ran our mixed-effects model to obtain a null distribution representing the decrease in semantic distance from a *random* speaker's utterance on the first round. We found that this permuted baseline also showed a linear decrease over time, but our true estimate ($b = -3.5$) fell narrowly outside the null distribution of effects (95% $CI = [-3.34, -3.04]$), showing that utterances by a particular speaker drifted from their own initial utterance to a slightly greater degree than would be expected due to generic differences between utterances made at different timepoints in an interaction [6]. This difference is likely a consequence of random utterances from different speakers being more dissimilar even on *early* repetitions, thus depressing the overall slope.

Increasing internal consistency within interaction

As speakers modified their utterances across successive repetitions, we additionally hypothesized that they would converge on increasingly consistent ways of referring to each tangram. To test this prediction, we computed the semantic similarity between successive utterances produced by each speaker when referring to same tangram (i.e. repetition k to $k + 1$). A mixed-effects model with linear and quadratic fixed effects of repetition number and maximal random effects for both tangram and pair of participants showed that similarity between successive utterances increased substantially throughout an interaction ($b = 2.7$, $t = 10.9$; Fig. 5B). The quadratic term was not significant ($b = -0.4$, $t = -1.8$). Again, we compared our empirical estimate of the magnitude of this trend to a null distribution of slopes estimated by scrambling utterances across pairs and re-running the regression model.The estimated slope fell outside this null distribution, for which similarity was strongly *decreasing*, $CI = [-5.9, -5.4]$, providing evidence

[6] Both here and for the permuted baselines in the subsequent two analyses we needed to simplify the random effects structure to contain only random intercepts due to convergence issues over the large number of permutations. However, we were interested in the coefficient estimate rather than statistical significance in these permuted models, and estimates appeared stable across different random effects structures.

that increasingly consistent ways of referring to each object manifested only for series of utterances produced within the same interaction.

Increasingly different content across interactions

Finally, we predicted that the way different pairs refer to the same tangram would become increasingly dissimilar from each other across repetitions, gradually diverging into different equilibria. We tested this prediction by computing the mean pairwise similarity between utterances used by different speakers to refer to the same object. The large sample of pairwise similarities ($N = 257,040 = 12\ \text{tangrams} \times 6\ \text{repetitions} \times \frac{85 \cdot 84}{2}\ \text{distinct pairs}$) presented both advantages and disadvantages. On one hand, we could obtain highly reliable estimates of mean similarity. On the other hand, larger random-effects structures led to convergence problems. We therefore ran a mixed-effects regression model including linear and quadratic fixed effects of repetition number including maximal random effects only at the tangram-level. We found a strong negative linear fixed effect of repetition on between-game semantic similarity ($b = -50.7, t = 16.8$) as well as a significant quadratic effect ($b = 16.1, t = 12$), indicating that this divergence slows over time as each pair stabilizes, (see Fig. 5C). We again conducted a permutation test to compare this t value with what would be expected from scrambling utterances across repetitions for each pair and target. We found that the estimated slope was highly unlikely under this distribution ($CI = -2.5, 2.9], p < 0.001$).

Visualizing trajectories through vector space

Finally, to better understand the changes uncovered by these analyses of utterance embeddings, we visualize the trajectories taken by each pair of participants when referring to a particular example tangram, annotating utterances in several parts of the space. First, we took the first 50 components recovered by running Principal Components Analysis (PCA) on the 300-dimensional utterance embeddings. We then use t-SNE (Maaten and Hinton, 2008) to stochastically embed the lower-dimensional PCA representation of each utterance in a common 2D vector space. In Fig. 6, each arrow connects the first and last utterance a particular pair used to refer to this tangram.

We observe that the initial utterances of each game tend to cluster tightly near the center of the space and the final utterances are *dispersed* more widely around the edges. This pattern is consistent with the hypothesis that different speakers overlap more in the content of their early descriptions before honing in on more distinctive different equilibria later in the game (see 3.1). For this particular tangram, there were a handful of semantically distinct labels that served as equilibria for multiple pairs ("ghost," "flying," "angel") as well as many more idiosyncratic labels. Pairs often initially mentioned multiple properties (e.g. "person raising their arms up like a choir singer") before breaking the symmetry and collapsing to one of these properties ("choir singer").

4 | RESULTS: CHARACTERIZING THE DYNAMICS OF STRUCTURE

In the previous section, we examined the dynamics of semantic content. We found that pairs converged systematically on distinctive words but different pairs discovered different solutions to the same coordination problem. Here, by contrast, we examine the dynamics of *structure*, testing the extent to which different pairs share more abstract commonalities in the structural changes of their referring expressions over time, even while differing in their content. In particular, we examine *how* different pairs reduce the length of their utterances. What sequence of transformations is applied to reduce long initial descriptions into shorter final ones?

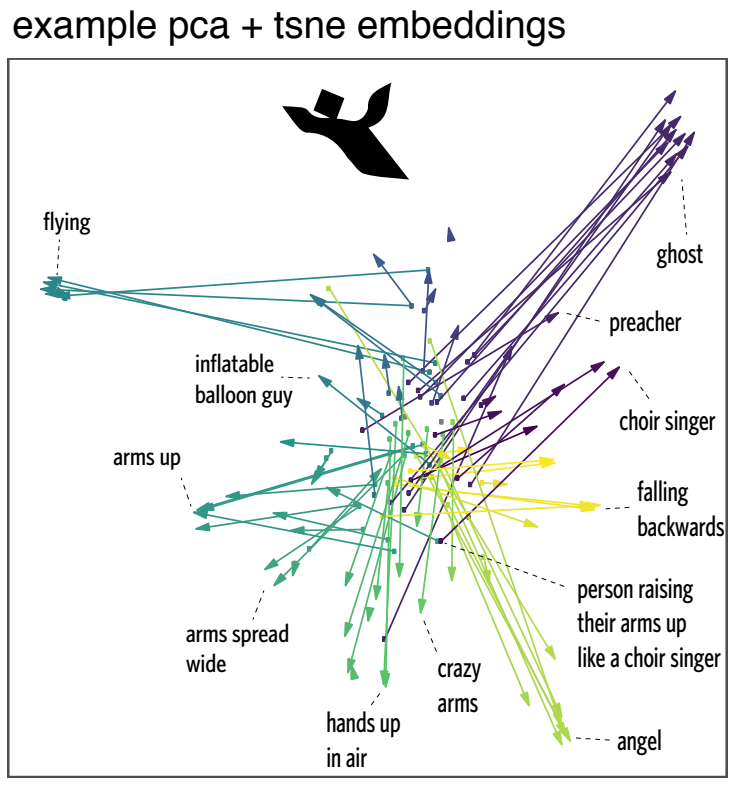

FIGURE 6 2D projection of semantic embeddings for example tangram. Each arrow represents the trajectory between the first round to last round for a distinct pair of participants. Color represents the rotational angle of the final location to more easily see where each pair began. Annotations are provided for select utterances, representing different equilibria found by different participants.

4.1 | Dialogue between speaker and listener

Before focusing our analysis on the dynamics of how *directors* transform their referring expressions over time, we first on the broader structure of bi-directional dialogue exchanges. Conceptual pacts are formed *collaboratively* (Clark and Wilkes-Gibbs, 1986): directors and matchers engage in a bi-directional process where matchers ask follow-up questions, suggest corrections, and acknowledge or verbally confirm their understanding through a backchannel. In the absence of feedback, descriptions may not necessarily get shorter (Krauss and Weinheimer, 1966; Garrod et al., 2007).

While we automatically supplied minimal feedback about the listener's response each round, we predict that the additional feedback from listener backchannel replies should be highest on the first round and drop off once meanings are agreed upon. To test this prediction, we coded whether the listener sent a message or not on each trial and fit a mixed-effects logistic regression model with a fixed effect of repetition, random intercepts and slopes for each pair of participants, and a random intercept for each target. We found that the probability of the listener sending a message decreased significantly over the game ($b = -0.84, t = -9.1, p < 0.001$; $Fig.7A$). In aggregate, 76% of listeners send at

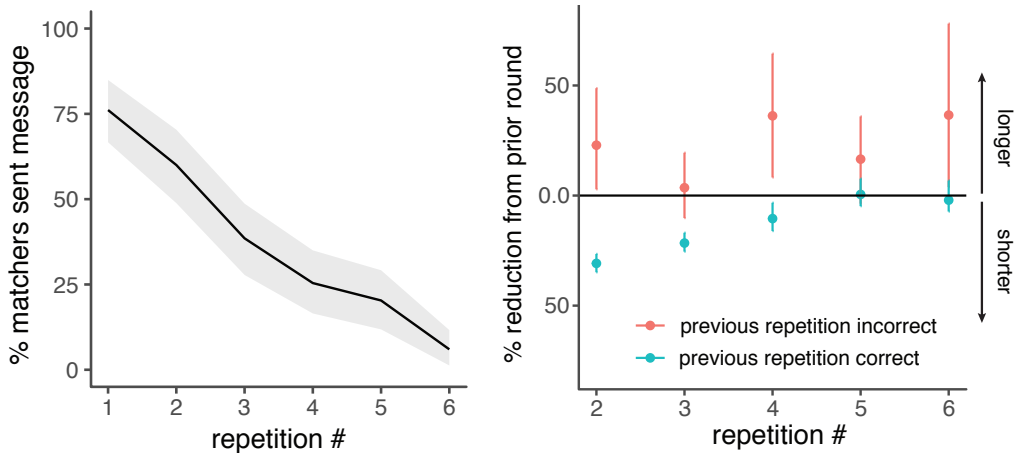

FIGURE 7 (A) Listeners reply with interactive feedback less often over repetitions, and (B) speakers are sensitive to listener response feedback, increasing message length on the subsequent repetition of a tangram after an error is made.

least one message in the first repetition block, but only 6% sent a message in the last block. These rates found in our online text-based replication are lower overall than in-person lab experiments, but we nonetheless strongly replicated the overall trend.

Next, we examined the extent to which speakers were sensitive to listener response feedback in modulating their utterances. If the listener failed to select the correct target, the speaker may take this as evidence that their description was insufficient and attempt to provide more detail the next time they must refer to the same tangram. If the listener is correct, on the other hand, the speaker may take this as evidence of understanding and reduce their level of detail on future repetitions. We tested these predictions by comparing the proportional change in utterance length on the repetition after an error against the change in length after a correct response (i.e. $(n_t - n_{t-1})/n_{t-1}$). This measure could be positive, indicating a net increase in utterance length, or negative, indicating a reduction.

We fit a mixed-effects regression model predicting this measure with an effect-coded categorical fixed effect of previous round feedback and a (centered) continuous effect of repetition number, including random intercepts and effects of feedback for each speaker. We found a significant main effect of feedback, even controlling for repetition: utterance length changed more in the shorter direction after correct responses than after negative responses, $b = -0.18, t = -6.2$ (see Fig. 7B). Indeed, speakers were more likely on average to *add* words on the repetition after an error at any point in the game. Because repetitions of the same tangram were spaced out and errors were relatively rare, this effect is unlikely to simply reflect heightened attention on trials after an error. Instead, this pattern of results is consistent with sensitivity to tangram-specific evidence of the listener's understanding.

4.2 | Understanding reduction

Next, we turn to a set of analyses examining the reduction in utterance length over the course of the experiment. At the coarsest level, we find that the mean number of words used by speakers decreases over time (see Fig. 8) in both the free matching and cued variants of the task. Free matching required more words overall because participants needed to additionally mention which tangram they were referring to (i.e. "number 3 is the …"). This result replicates

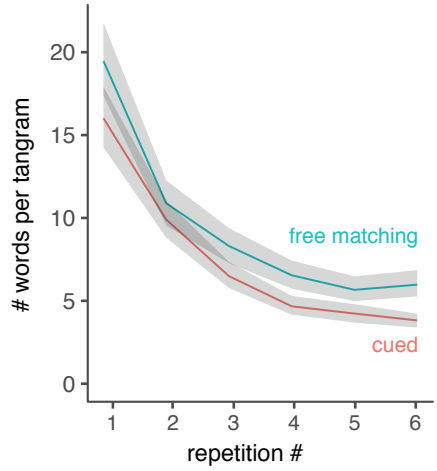

FIGURE 8 Similar reduction in # words per tangram for both variants of the task. Error ribbons are 95% confidence intervals.

the highly reliable reduction effect found throughout the literature on repeated reference games (e.g. Krauss and Weinheimer, 1964; Brennan and Clark, 1996). Perhaps because of the text-based (vs. spoken) interface, participants in our task used fewer words overall than reported by Clark and Wilkes-Gibbs (1986). The following analyses break down this broad reduction into a finer-grained set of phenomena on the *cued* corpus.

Reduction in parts of speech

The first level of granularity concerns which *kinds* of words are most likely to be dropped. We used the SpaCy part-of-speech tagger (Honnibal and Montani, 2019) to count the number of words belonging to different parts of speech in each message. In Fig. 9A, we show the shifting proportions of different parts of speech at each repetition, We find that nouns account for proportionally more of the words being used over time, while determiners and prepositions account for fewer. To test which kinds of words are more likely to be dropped, we measured the percent reduction in the number of words in each part of speech from the first round to the sixth round. We find that determiners ('the', 'a', 'an') are the most likely class of words to be dropped (90%) and nouns ('dancer', 'rabbit') are the least likely to be dropped (61%).More generally, closed-class parts of speech, including function words like determiners, are strictly more likely to be dropped than open-class parts of speech (Fig. 9B). Because open-class parts of speech are statistically more likely to supply *distinctive* words than closed-class parts of speech, these structural considerations may contribute to the patterns in distinctiveness reported in section 3.1.

One possible interpretation of these findings is that reduction may be driven mostly by the loss of function words as speakers shift to a less grammatical shorthand over the course of the task. However, when examining the n-grams most likely to be dropped (see Table 1), we find that many of the most dropped closed-class words are used to form conjunctions ('and') or prepositional phases ('of', 'with'). Others are modifiers ('the right …'). These examples suggest an alternative explanation: the higher reduction of closed-class function words may be a consequence of entire meaningful grammatical units being dropped at once. If initial descriptions tend to be syntactically complex, combining multiple partially redundant sources of information for identifying the target, then the speaker may omit

	unigrams	bigrams	trigrams
#1	a	look like	look like a
#2	the	like a	look like -PRON-
#3	-PRON-	to the	to the right
#4	like	-PRON- be	like -PRON- be
#5	be	this one	like a person
#6	look	the right	to the left
#7	on	the left	this one look
#8	one	like -PRON-	one look like
#9	with	on the	this one be
#10	to	with a	-PRON- look like
#11	and	a person	look like someone
#12	right	on top	diamond on top
#13	this	in the	in the air
#14	of	a diamond	on top of
#15	head	have a	a diamond on

TABLE 1 Top 15 unigrams, bigrams, and trigrams with the highest numeric reduction from first round to last round. Text lemmatized before n-grams computed.

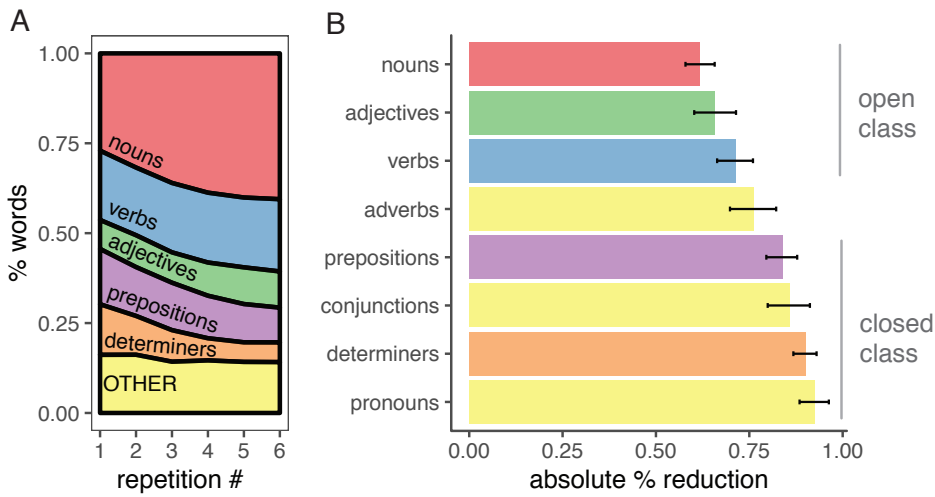

FIGURE 9 (A) proportion of words in different part of speech on each repetition. (B) Closed-class parts of speech are more likely to be dropped than open-class parts of speech. Error bars are bootstrapped 95% confidence intervals.

entire modifying clauses with additional evidence of a listener's understanding.

Reduction in syntactic constituents

We explicitly examined this hypothesis by examining whether dropped words tend to come from the same syntactic units, relative to a random deletion baseline. We quantified the extend to which dropped words 'cluster' by examining

FIGURE 10 Example dependency parse for referring expression. If the sub-phase "arms out in front" were dropped, we would find a mean dependency length of 1.33 among the dropped words.

dependency lengths between the dropped words (Jurafsky and Martin, 2014; Futrell et al., 2015). Specifically, we compared each referring expression to the one produced on the subsequent repetition to determine which words were dropped. Then we looked up each pair of dropped words and found the shortest path between them in the dependency parse tree (see Fig. 10). We then took the mean dependency length across all such pairs in all games.

This empirical 'syntactic clustering' statistic was then compared to a random baseline. For this baseline, instead of examining dependency lengths for the words that were actually dropped, we randomly sampled the same number of words from the referring expression and computed the dependency length between them. We repeated this procedure 100 times to obtain a null distribution of the mean dependency length that would be expected if words were being dropped *randomly* from anywhere in the message.

We found a mean empirical dependency length of 4.22, which lay outside the null distribution ($95\% CI$: [4.38, 4.44]), indicating a small but reliable effect of syntactic clustering. The words that were dropped tended to be closer to one another in the dependency parse than expected by chance. Other statistics such as the minimum dependency path or the raw distance in the sequence of words gave similar results. This result accords with early observations by Carroll (1980), which found that in three-quarters of transcripts from Krauss and Weinheimer (1964) the short names that participants converged upon were prominent in some syntactic construction at the beginning, often as a head noun that was initially modified or qualified by other information.

5 | DISCUSSION

In this paper, we characterized the dynamics of convention formation in a classic repeated reference game paradigm. First, we found that pairs of participants systematically tend to conventionalize words that are more distinctive in the initial context. Using both discrete and continuous measures of semantic content, we found this process leads to stable usage within pairs but multiple equilibria across different pairs. Second, speakers reduce the length of their descriptions over the course of the task in a way that is sensitive to evidence of listener understanding and syntactically structured to omit entire meaningful units. In sum, these results establish new benchmark phenomena that computational models of adaptation and convention formation must account for.

Discrete vs. continuous semantic measures

Because our results hinged on our quantification of semantic content, it is worth noting some advantages and disadvantages of discrete measures compared to continuous vector space measures. A key advantage of measures based on the word distribution is that the entropy is not dependent on any particular choice of pre-trained vector embedding. Due to biases in the training corpora, vector representations also may not capture some of the more idiosyncratic conventions that participants converge on (e.g. "YMCA" or "zig zag" or "Frank" – short for "Frankenstein"). To the extent we find converging results, the discrete measure may address concerns about the quality of the continuous

representation.

A key disadvantage is that the entropy is sensitive to the support of the word distribution — the vocabulary present in the corpus on a particular round — and thus does not have a natural scale. While directly measuring divergence between word distributions at different repetitions and between different pairs is technically possible, their sparsity makes this approach not as informative at these finer granularities of analysis. Many pairs use entirely disjoint sets of words, and on later rounds, the distribution may only contain one or two words. Further, because it is based entirely on the frequency of tokens, it may treat even close synonyms as entirely distinct tokens in the word distribution. Thus, these two approaches provide complementary evidence for the dynamics of content.

Symmetry-breaking

Our results in this paper also raise a subtle cognitive question about classic definitional notions of arbitrariness in convention (Lewis, 1969), which hold that there must counterfactually exist an alternative solution to the coordination problem for any particular solution to be conventional. How can such systematicity in the formation process co-exist with conventionality? The symmetry of different solutions must break somewhere to account for the empirical existence of many alternative but equally successful referential conventions at the population level, but our results are consistent with two different cognitive realities *within* games.

One possibility is that this arbitrariness is a product of substantial variability in a population of fairly rigid speakers. That is, each individual speaker may have strong but idiosyncratic initial preferences for how to refer to each tangram. They may begin with additional elaboration given their representation of uncertainty about whether these strong preferences are shared, and in the absence of misunderstandings they will persist with their preferred and premeditated label. A second possibility is that the population of speakers is homogenous but only weakly constrained by prior preferences. That is, speakers may not only be uncertain of which messages their partner will understand, but are themselves unclear on an appropriate way to refer to these unfamiliar objects. If speakers initially produce an utterance from a broad distribution of acceptable labels, and update their distribution on subsequent rounds, different pairs may end up in different equilibria due to the path-dependence of the process. In this way, the symmetry may be broken through randomness in the sampling step (or in the updating step, if learning is stochastic.) These possibilities are not mutually exclusive. Our results rule out the possibility of a homogeneously rigid population, but it is possible that some speakers have strong preferences about labels while others are uncertain.

While prior empirical work has indirectly tested initial expectations and preferences – for instance, by asking speakers to either produce descriptions for others or for themselves in the future (Fussell and Krauss, 1989) – an important direction for further work is to design experiments that disentangle these possibilities. For instance, a Bayesian truth serum approach (Prelec, 2004) could estimate both an individual's own subjective preferences and their expectations about whether these would be shared by others. More broadly, following recent attempts to estimate the true variability across speakers in phonetic properties of speech production (Kleinschmidt, 2019), it would be valuable to estimate how much variability in semantic expectations there really is in the population (Furnas et al., 1987).

references

Arora, S., Liang, Y. and Ma, T. (2017) A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations (ICLR)*.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003) A neural probabilistic language model. *Journal of machine learning research*, **3**, 1137–1155.

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

Brennan, S. E. and Hanna, J. E. (2009) Partner-specific adaptation in dialog. *Topics in Cognitive Science*, **1**.

Carroll, J. M. (1980) Naming and describing in social communication. *Language and Speech*, **23**, 309–322.

Clark, H. H. (1983) Making sense of nonce sense. *The process of language understanding*, 297–331.

— (1998) Communal lexicons. In *Context in language learning and language understanding* (eds. K. Malmkjaer and J. Williams), chap. 4, 63–87. Cambridge: Cambridge University Press.

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

Davidson, D. (1986) A nice derangement of epitaphs. *Philosophical grounds of rationality: Intentions, categories, ends*, **4**, 157–174.

Delaney-Busch, N., Morgan, E., Lau, E. and Kuperberg, G. R. (2019) Neural evidence for bayesian trial-by-trial adaptation on the n400 during semantic priming. *Cognition*, **187**, 10–20.

Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T. (1987) The vocabulary problem in human-system communication. *Communications of the ACM*, **30**, 964–971.

Fussell, S. R. and Krauss, R. M. (1989) The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, **25**, 203–219.

Futrell, R., Mahowald, K. and Gibson, E. (2015) Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, **112**, 10336–10341.

Garrod, S., Fay, N., Lee, J., Oberlander, J. and MacLeod, T. (2007) Foundations of representation: where might graphical symbol systems come from? *Cognitive Science*, **31**, 961–987.

Glucksberg, S. and McGlone, M. S. (2001) *Understanding figurative language: From metaphor to idioms*. No. 36. Oxford University Press on Demand.

Goodman, N. D. and Frank, M. C. (2016) Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, **20**, 818 – 829.

Hawkins, R. X. D. (2015) Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, **47**, 966–976.

Honnibal, M. and Montani, I. (2019) spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Hupet, M. and Chantraine, Y. (1992) Changes in repeated references: Collaboration or repetition effects? *Journal of psycholinguistic research*, **21**, 485–496.

Jurafsky, D. and Martin, J. H. (2014) *Speech and language processing*, vol. 3. Pearson London.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A. and Fidler, S. (2015) Skip-thought vectors. In *Advances in neural information processing systems*, 3294–3302.

Kleinschmidt, D. F. (2019) Structure in talker variability: How much is there and how much can it help? *Language, cognition and neuroscience*, **34**, 43–68.

Krauss, R. M., Garlock, C. M., Bricker, P. D. and McMahon, L. E. (1977) The role of audible and visible back-channel responses in interpersonal communication. *Journal of personality and social psychology*, **35**, 523.

Krauss, R. M. and Weinheimer, S. (1964) Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, **1**, 113–114.

— (1966) Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, **4**, 343.

Landauer, T. K. and Dumais, S. T. (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, **104**, 211.

Lascarides, A. and Copestake, A. (1998) Pragmatics and word meaning. *Journal of linguistics*, **34**, 387–414.

Lewis, D. (1969) *Convention: A philosophical study*. Harvard University Press.

Maaten, L. v. d. and Hinton, G. (2008) Visualizing data using t-sne. *Journal of machine learning research*, **9**, 2579–2605.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013) Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Osgood, C. E. (1952) The nature and measurement of meaning. *Psychological bulletin*, **49**, 197.

Pennington, J., Socher, R. and Manning, C. (2014) Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Piantadosi, S. T., Tily, H. and Gibson, E. (2012) The communicative function of ambiguity in language. *Cognition*, **122**, 280–291.

Pickering, M. J. and Garrod, S. (2004) Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, **27**, 169–190.

Prelec, D. (2004) A bayesian truth serum for subjective data. *science*, **306**, 462–466.

Robertson, S. (2004) Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, **60**, 503–520.

Salton, G. and Buckley, C. (1988) Term-weighting approaches in automatic text retrieval. *Information processing & management*, **24**, 513–523.

Sparck Jones, K. (1972) A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, **28**, 11–21.

Wilkes-Gibbs, D. and Clark, H. H. (1992) Coordinating beliefs in conversation. *Journal of memory and language*, **31**, 183–194.