Journal of Language Evolution, 2019, 1–18
doi: 10.1093/jole/lzy010
Advance Access Publication Date: 30 October 2018
Research article

Social Group Effects on the Emergence of Communicative Conventions and Language Complexity

Mark Atkinson [1,*] Gregory J. Mills[2,3] and Kenny Smith[4]

[1]Department of Psychology, University of Stirling, Stirling FK9 4LA, UK, [2]Center for Language and Cognition, University of Groningen, 9712 CP Groningen, Netherlands, [3]School of Informatics, The University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB and [4]School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, 3 Charles Street, Edinburgh EH8 9AD, UK

*Corresponding author: E-mail: mark.atkinson@stir.ac.uk

Abstract

Languages differ in their complexity. One possible explanation for this observation is that differences in social factors influence linguistic complexity: languages that are used for communication in small-scale 'societies of intimates' exhibit greater complexity as a result of the communicative contexts in which they are typically employed. We used the techniques from referential communication studies across three experiments to assess the effects of two social group factors—group size and amount of communally shared knowledge—on the brevity and transparency of linguistic conventions. In Experiment 1, we explored the effects of a manipulation of group size, comparing the conventions which develop from the interaction of two speakers, with those which develop between three speakers. In Experiment 2, we manipulated the extent to which groups of three speakers share talk-relevant contextual information. While we found the conditions that involve larger groups and less shared background information initially resulted in longer labels and a greater reliance on more literal descriptive terms, there was no effect of either factor in the longer term. In Experiment 3, we investigated the transparency of the conventions of Experiments 1 and 2 by assessing how well they could be matched to their intended referents by naive individuals. We found no evidence to support the claims that communicative contexts involving communicating with more individuals, or individuals with whom less relevant information is shared, produce more transparent conventions. Our experiments ultimately provide no support for the idea that the structure of linguistic conventions is shaped by the groups in which they develop.

Key words: interaction; linguistic conventions; language complexity; social group effects; esoteric communication

1. Introduction

Languages are shaped by learning and use (Kirby 1999; Croft 2000; Christiansen and Chater 2008; Smith and Kirby 2008; Beckner et al. 2009). Since the pressures from learning and use are likely to be different in different types of social group and in different social contexts, it has been claimed that non-linguistic factors may systematically influence the characteristics of languages (Croft 1995; Nettle 1999; Wray and Grace 2007; Trudgill 2011; Dale and Lupyan 2012). In this paper,

we investigate the role that group size and the amount of communally shared knowledge may have on the form of new communicative conventions, adapting techniques from referential communication experiments to test whether these social factors shape emerging communicative conventions. In doing so, we aim to both extend the literature of referential communication to consider the effects of group size and the amount of knowledge shared by interlocutors, and to see how the paradigms used in referential communication studies can be adapted to investigate how social structure shapes language.

1.1 Social structure shapes language structure

A number of theories connect sociocultural factors to structural properties of language (e.g., Wray and Grace 2007; Trudgill 2011), resulting in cross-linguistic variation in language transparency and complexity. By complexity, we refer to the descriptive complexity of a language here, considering complexity as an inherent and objective property of a linguistic system. Although all languages may be viewed as equally complex in that they have the potential to combine a finite number of elements to convey infinitely many possible meanings, this does not mean that the encoding of meaning in signals is necessarily equally complex (Nettle 2012), and the complexity of that encoding may, at least partly, be influenced by the environment—sociocultural or otherwise—in which an individual language is learned and used (Lupyan and Dale 2016).

Wray and Grace (2007) consider two extreme social contexts for communication, *esoteric* and *exoteric*, and the potential impact of these social contexts on language structure. Esoteric communication occurs within small groups, with relatively simple social network structures, in which speakers can rely on a large amount of shared knowledge and experience with their interlocutors, and where contact with strangers and other languages is limited; a so-called 'society of intimates' (Givón 1979: 297). Exoteric communication occurs in larger groups with more complex social networks, in which shared knowledge and experience is more limited. According to theories linking social structure and linguistic structure, the languages of groups where esoteric communication is the norm are structurally more complex, have more irregular forms, and have less transparent form-meaning mappings, where semantic categories map less predictably to linguistic expressions (Wray and Grace 2007; Lupyan and Dale 2010; Trudgill 2011); they may also have greater levels of syntagmatic and paradigmatic redundancy (Lupyan and Dale 2010), and a greater number of more semantically specific lexical items (Wray and Grace 2007). By contrast, the languages of groups where exoteric communication is widespread have simpler, more regular, grammars with more transparent compositional structure, being consequently easier for out-group members to understand and learn. Analyses of large datasets at least partially support these claims, suggesting that languages with greater numbers of speakers have lower levels of grammatical complexity (Nichols 2009; Sinnemäki 2009; Lupyan and Dale 2010).

One prominent theory explaining the link between population structure and linguistic complexity is that languages which have a larger number of speakers are simpler due to the effects of adult learning (Wray and Grace 2007; Lupyan and Dale 2010; Trudgill 2011; Nettle 2012; Atkinson et al. 2015). Languages with more speakers are also typically those with a greater proportion of non-native speakers (Lupyan and Dale 2010), and there is evidence that adult learners find particular linguistic features, such as morphological complexity, irregularities, and syntagmatic and paradigmatic redundancy, particular challenging to acquire (Wray and Grace 2007; Clahsen et al. 2010; Lupyan and Dale 2010; Trudgill 2011; Lupyan and Dale 2016). Languages with greater degrees of adult contact and learning might therefore adapt to the needs and abilities of adult learners, with the languages features which are specifically challenging for adults to acquire filtered out (Wray and Grace 2007; Lupyan and Dale 2010; Bentz and Winter 2013).

An alternative or complementary account, which we explore here, is that differences in linguistic structure which correlate with social structure might be a result of differences in language *use* and communicative context, rather than differences in language learning. If individuals are more likely to share interests, occupations, cultural practices, and experiences—there is 'shared knowledge' (Wray and Grace 2007), 'communally shared information', or 'informational homogeneity' (Trudgill 2011)—their communicative needs and preferences are likely to be different than if they share less (Sapir 1912; Wray and Grace 2007; Trudgill 2011). More communally shared information, argue Wray and Grace (2007) and Trudgill (2011), will lead to a greater likelihood that interlocutors will share specialized vocabulary and be better able to exploit pragmatic context. In the interests of processing efficiency, more specific, or semantically more complex, lexical items are then more likely to be employed. Conversely, if there is less communally shared knowledge, there is a greater potential for errors in hearer comprehension; speakers

may have to employ more common lexical items to increase the chance that they share them with hearers, and encode their signals in a more systematic way to allow hearers to determine meaning from their composition.

This theory linking communicative context to linguistic complexity has received little in the way of direct experimental tests. However, as we review below, the hypothesis that communicative context can affect efficiency and comprehensibility has been explored in a series of referential communication studies.

1.2 Experimental studies of referential communication

We use techniques established in the study of naturalistic dialogue to explore this potential link between group size, shared knowledge, and complexity in communication. Experimental studies of dialogue and the emergence of communicative conventions go back to Krauss and Weinheimer (1964) and their investigation into the development of referring expressions. Krauss and Weinheimer (1964) had pairs of participants repeatedly describe novel images in English; the more times an image was encountered and described, the shorter its description became. One pair, for example, initially described an image as 'upside-down martini glass in a wire stand'. With repeated interaction, this reduced to 'inverted martini glass', then 'martini glass', and finally 'martini'.

Clark and Wilkes-Gibbs (1986) argue that these conventionalized referring expressions emerge through a process of collaboration, with both the speaker and the hearer involved in establishing successful communication. By this account, a description becomes grounded in that it is proposed by the speaker, and refined by either or both interlocutors until it is accepted by both parties. A potentially idiosyncratic description produced by one speaker is therefore developed until it is mutually understood. Once such an expression has entered the dyad's common ground, it reduces in length as the speakers increase the efficiency of their interaction.

Other studies support this collaborative view. In a study by Hupet and Chantraine (1992), participants were required to repeatedly label sets of tangrams, and told that their descriptions would either be given to the same recipient for each repetition, or to a different recipient each time. The descriptions did not reduce in either case, suggesting that mere repetition is not sufficient: mutual acceptance of a description is necessary for it to become shorter. If feedback is also given while a referent is being described, as opposed to only after a description is completed, the referring expressions shorten even more rapidly (Krauss and Weinheimer 1966).

Intended audience, interaction, and being actively involved in the negotiation process also influence how easy referring expressions are to comprehend. Fussell and Krauss (1989) found that descriptions written for other people are longer and more literal than personal ones designed for the writer themselves, and that they were easier to match to their intended referents by a naive individual. A speaker may refer to an image as 'a rectangle with a series of curves attached to it by diagonal lines' when it is intended for another person, for example, but eschew geometric terms and use the more figurative 'spider' for themselves. Monologue descriptions, even when intended for others, are also more difficult to comprehend than those arising through dialogue (Fox Tree 1999). This may be because dialogues contain a greater number of perspectives, and so increase the likelihood of there being a perspective which is understood by a third person (Fox Tree and Mayer 2008), or the grounding process may increase the likelihood that the descriptions will be comprehendable by any individual, not just those directly involved in the interaction (Branigan et al. 2011). Overhearers (who observe an unfolding dialogue but do not participate in it) are also less accurate at identifying referents from descriptions than those involved in the negotiation themselves: being present throughout the process does not give the same advantage in comprehension (Schober and Clark 1989), probably because the overhearer cannot guide the developing description to one which they would prefer to adopt (Branigan et al. 2011). Speakers, however, are sensitive to potential comprehension limitations of interlocutors who have not played a part in the negotiation process, and may compensate by using longer descriptions (Yoon and Brown-Schmidt 2014), even if increasing the number of speakers who played no part in the negotiation process may not increase the length of those descriptions further (Rogers et al. 2013).

Similar techniques have been extended to the development of non-linguistic, graphical communication studies (see Galantucci and Roberts [2012] for review). In a classic study, Garrod et al. (2007) demonstrated the importance of interaction on the development of arbitrary symbols from iconic images, and showed that individuals not involved in the grounding process were less able to correctly interpret the resultant signs. Subsequent studies have shown that similar processes operate in larger communities: completely shared knowledge of the grounding process across all members of a population is not necessary, and simple graphical symbols can emerge even with population turnover (Fay et al. 2008, 2010; Caldwell and

Smith 2012). Intriguingly however, signs emerging in groups are more transparent (i.e., their meaning can be more easily guessed by naive individuals) than those which emerge in dyads, even though they are equally reduced in form and do not differ in their complexity; in both cases, the signs are initially iconic, but with repeated use those in the group condition simplify while retaining iconic properties which allow them to be easily interpreted (Fay et al. 2008, 2010).

1.3 The present study

In the following experiments, we investigate how the emergence of linguistic conventions is affected by social group size and the contexts in which group members communicate. In doing so, we aim to assess the claims that some communicative contexts will produce more complex, less transparent language use than others. In Experiment 1, we extend the experimental method from Clark and Wilkes-Gibbs (1986) to compare the referring expressions which emerge in dyads and triads (groups of three interlocutors), assessing description lengths, transparency, and semantic complexity. Although it has been proposed that group size alone may not influence language features (Lupyan and Dale 2010; Nettle 2012), it is nevertheless one of the features proposed to distinguish more esoteric and more exoteric communities and communicative contexts (Wray and Grace 2007; Trudgill 2011) and, as discussed above, the referential communication literature has shown that the presence of just a third speaker may reduce comprehension at group level and elicit longer descriptions. If we see differences in communicative conventions even when increasing the size of group from two to just three speakers, then we may anticipate that group size itself may have some effect in more naturalistic contexts when of course differences in the number of speakers will be a lot more pronounced. As we will see below, our group size manipulation does lead to quantifiable differences between the initial descriptions produced in each condition. In Experiment 2, we compare the triadic condition of Experiment 1 to a second triadic condition where we reduce the amount of talk-relevant information—one possible means of reducing the 'communally shared information' (discussed above)—shared by the three members of the group. Although we recognize that shared knowledge is but one characteristic separating more esoteric and exoteric communicative contexts, reducing shared knowledge while keeping the other features of the group constant would still reduce esotericity.

We expect that repeated interaction will result in shorter description lengths (as has been shown repeatedly for dyads, cf., e.g., Krauss and Weinheimer 1966) in all cases. We then consider the effect of esotericity on linguistic complexity. While languages can differ in complexity at multiple levels (e.g., morphosyntactic, phonemic), here we focus on specific claims in the literature regarding the effects of esotericity which can be studied using natural language referential communication paradigms. Specifically, across these two experiments, we test whether smaller group size or more shared knowledge results in shorter descriptions, fewer literal descriptive terms (cf., Fussell and Krauss 1989), less transparent form-meaning mapping between the referents and the labels participants use to describe them (Wray and Grace 2007; Lupyan and Dale 2010; Trudgill 2011), and more semantically complex lexical items (Wray and Grace 2007; Trudgill 2011).

Finally, in Experiment 3, we assess the descriptions from Experiments 1 and 2 for transparency, by asking naive individuals to match them to their intended referents. We investigate whether those produced in larger groups, or by interlocutors with less shared information, are easier to identify (Wray and Grace 2007; Fay et al. 2008, 2010; Fox Tree et al. 2008; Branigan et al. 2011; Trudgill 2011).

2. Experiment 1: the effect of group size

Participants played a communication game in a small group of two or three participants: in the Dyad condition, two participants completed the experiment together; in the Triad condition, participants completed the experiment in groups of three. In both conditions the group's task was to describe tangrams (abstract geometrical shapes) for the other participant(s) in their group, and to select tangrams from a larger set based on the descriptions provided by the other member(s) of their group. Each group played multiple such rounds of communication, repeatedly describing the same tangrams. Although we are not suggesting that a group of three speakers should be considered an exoteric community in a naturalistic context, nor that group size in itself is necessarily the most important distinction between esoteric and exoteric communities, the Dyad condition can still be seen as a relatively esoteric communicative context due to the lower number of speakers.

2.1 Materials and methods

2.1.1 Participants

Sixty-two participants (forty-one female, twenty-one male; aged between 18 and 40 years, mean 21.3) were

recruited via the Student and Graduate Employment Service at the University of Edinburgh. They were recruited either individually and placed with other participants in a dyad or triad, recruited as a pair and placed with a third participant to make up a triad, or else they signed up in groups of two or three to participate as self-selected dyads or triads. Twenty-four participants were assigned to the Dyad condition; thirty-six participants were assigned to the Triad condition. Participants in the Triad condition were paid £7 for around 60 min; in the Dyad condition £5.50 for around 45 min. Data from fifty participants (ten dyads, ten triads) were retained, the remaining participants' data being discarded for failure to understand the task after repeated instruction (as indicated by continued discussion and uncertainty over the experimental task; two participants in a single dyad) or failure to complete six rounds in the allotted time (ten participants total; two dyads and two triads).

2.1.2 Materials

We constructed a set of forty-eight tangrams (see Fig. 1 for examples), made up of four sets of twelve (subjectively) related tangrams: 'animals', 'birds', 'people', and 'trinkets'.[1] For each group of participants, twelve tangrams were randomly selected from this larger set as target images, those which would be the targets for description during the experiment. Twelve additional images were randomly selected for each group to act as foils, which were never a target for description but which could be (erroneously) selected by participants when attempting to identify which tangram was being described by their partner(s). There was no stipulation that either the targets or the foils had to be composed of equal numbers from each set.

2.1.3 Procedure

The experiment was run using the Dialogue Experimental Toolkit.[2] Participants played together in their group, describing and matching tangrams over a number of rounds. At the start of each round, each participant was presented with a 6×4 array displaying the twenty-four tangrams (twelve potential targets plus twelve foils), presented in a random, participant-specific, configuration. They communicated with each other via the interface provided by the Dialogue Experimental Toolkit, which includes an instant-messaging chat window—participants simply typed text into the chat window, then hit return, at which point the message appeared in the chat window of all participants in the group. Message sender was indicated by the sender's username (selected by the participant), with the last few lines of the dialogue visible to all participants.

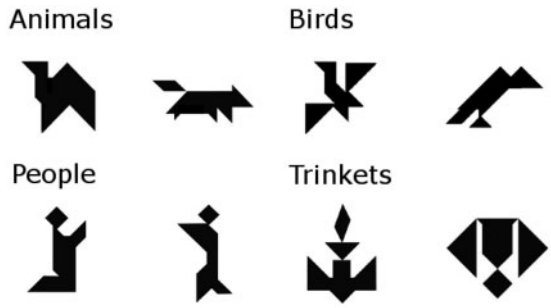

Figure 1. Example images from the set of forty-eight tangrams; two from each of the sets of Animals, Birds, People, and Trinkets. These sets are based on the tangrams' (subjective) similarity.

For a single round in the Dyad condition, eight target images were randomly selected from the larger set of the twelve potential targets (with a fresh selection being made on each round). Four of these images were assigned to each participant (the *director* for those images) to describe to their partner (the *matcher* for those images), and these images were marked with a blue border on the director's screen. Participants were able to select (and subsequently deselect) any of the other tangrams in their grid (i.e., those not marked with a blue border) using the mouse. Selected tangrams were marked with an orange border. The tangrams could be directed and matched in any order, that is, there was no requirement for the participants to alternate between director and matcher roles, nor for one participant to describe all of the tangrams they were assigned to direct in one go, etc. Figure 2 illustrates the experimental set up for a single participant near the start of a round. When both participants had selected exactly four tangrams (those which they believed were being described by their partner), either participant could end the round. Feedback was then given on the directed and selected tangrams (Fig. 3).

The Triad condition followed the same procedure, but at each round nine of the twelve target images were selected, and each participant was assigned three images to describe to the other group members, with the aim being for each individual to correctly select the six tangrams being described by their two partners. The chat windows displayed the messages for all three participants, with each message sender indicated by the sender's username as in the Dyad condition. All participants were able to interact with each other at all times, that is, there was nothing to prevent the two matchers from

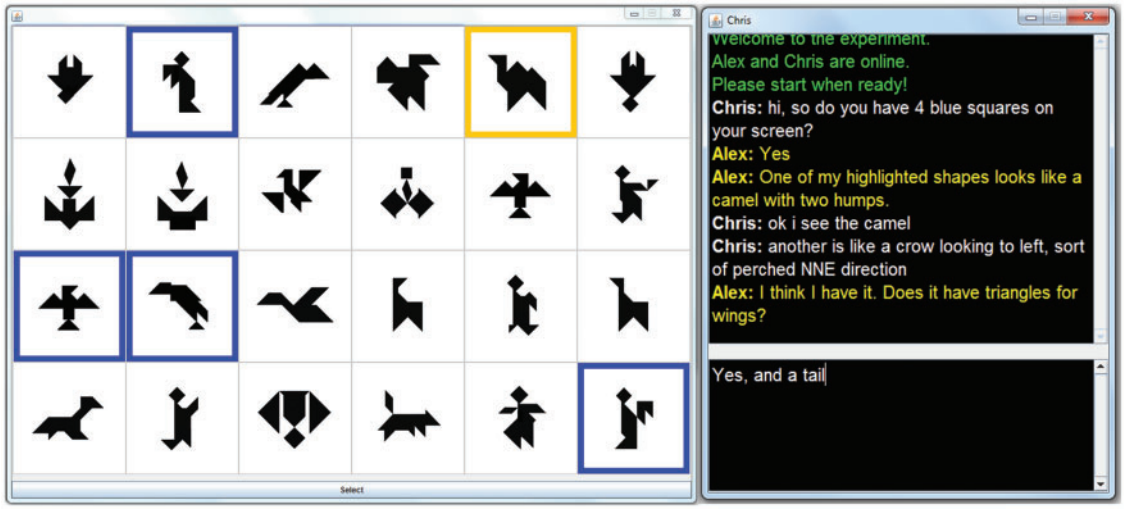

Figure 2. Example screen for a single participant in the Dyad condition at the start of a round. The images the participant has to direct to their partner are marked by blue borders. The orange border indicates the participant has selected an image they believe (in this case, correctly) has been described by their partner, in this case the one which their partner described as 'looks like a camel with two humps'. In the Triad condition, each participant would have three images to direct, and so three images marked with a blue border, while the messages of all three group members would be visible in the chat window.

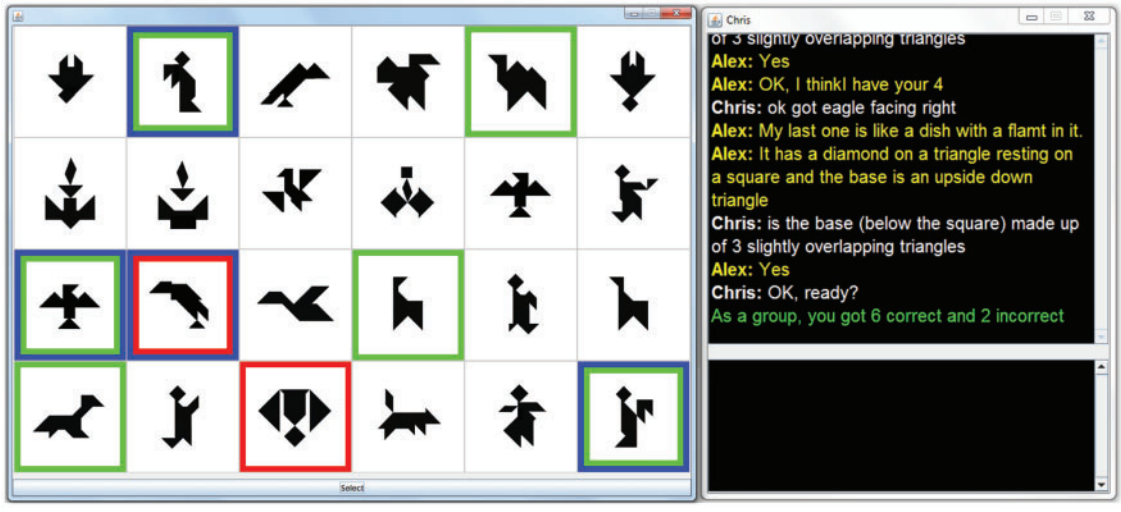

Figure 3. Example end of round feedback screen in the Dyad condition. Green and red borders indicate correct and incorrect selections, respectively. Those within the blue borders indicate the images which the participant's partner had (in)correctly matched—for triads, a red border indicated that their partner (or, in the Triad condition, at least one of their two partners) had mismatched. Those without blue borders are the participant's selections. In this case, the participant has incorrectly selected one of the four images directed by their partner, and their partner has incorrectly matched one which they directed.

interacting with each other while the third participant was directing.

In both conditions, we aimed to collect a minimum of six rounds of data, and groups who failed to reach this minimum were excluded from analysis (see exclusion information above).

2.2 Extracting descriptions of tangrams

As a consequence of the participants being able to freely interact using the chat window and describe the images in any way and in any order, the descriptions themselves were surrounded by a large quantity of additional linguistic material.[3] We therefore adopted the following

process to isolate the text we were interested in for analysis purposes—the actual descriptions produced by the director for each image—from the surrounding dialogue. First, we isolated all *director* lines of text, which we defined as those used by participants to describe the image they had highlighted in blue on their screen. These expressions included responses to matcher questions, including simple confirmations (e.g., 'yes the giraffe'). Secondly, we trimmed the director lines, preserving only the text which directly described the images. So, for example, 'i got the other giraffe' was trimmed to 'the other giraffe',[4] 'no' and 'that's the one' (responses to descriptions given by the matchers) were removed, and 'looks like' and 'my last one's like' were reduced to 'like'. Markers of certainty or reference to descriptions in previous rounds were retained.

Finally, and to preserve participant anonymity, each character in a participant name or username which was part of a description was replaced with 'X', resulting in, for example, 'XXXXX's big bird looking to the sky'. The trimmed director lines for each image were then concatenated to make what we consider the 'description' for the purposes of analysis. As an example, one triad description for an image in Round 1 was 'like a fox with a little tail that is howling like a wolf upwards to the right'; the same image was described as 'howling wolf' in Round 6.[5]

Two of our analyses below (the measures of systematicity and semantic specificity) also require that we identify the head word of each description. To isolate the head of a description, we first isolated the grammatical head of the main (i.e., most informative for descriptive purposes) phrase. As it was common and uninformative, the word 'one' was ignored; for example, in the phrase 'animal one', we took the head to be 'animal'. Where two words could be identified as the head, the first word was taken. For example, in 'like an emu or ostrich …', the head was taken to be 'emu'. Plurals were singularized where the description had originally referred to multiple images (e.g., 'men' was coded as 'man'), but not where the plurality was part of the description of a single image (e.g., 'triangles').[6]

As an example of the process, consider the following Round 1 exchange from one of the dyads:

Director: ok. do you have a fox?
Matcher: not sure what looks like a fox to you
Director: it's horizontal. with a triangular head on the right side. three legs. and a long rhombus shaped tail on the left
Matcher: gotya. and the whole shape is kind of together, only the tail is like standing up and barely connected to the rest of the shape, correct?
Director: yep

The concatenated director lines are then:

ok. do you have a fox? it's horizontal. with a triangular head on the right side. three legs. and a long rhombus shaped tail on the left yep

Which we trim to give us the following description for analysis purposes:

a fox? it's horizontal. with a triangular head on the right side. three legs. and a long rhombus shaped tail on the left

We then take 'fox' as the head for this description.

2.3 Dependent variables

We analysed five dependent variables to track the evolution of description schemes, which captured the functionality (two measures: communicative success and description length), semantic specificity (one measure), and transparency (two measures: use of geometric descriptions and systematicity of mapping) of the evolving descriptive conventions.

We analysed the communicative success scores for each condition by round. For the other analyses, we instead considered the descriptions grouped by occurrence: an Occurrence 1 description was the first time a given image was selected for description, regardless of whether or not that occurred in Round 1. We considered only the first four occurrences of a given image and its descriptions, as the number of images described five or six times was too low for meaningful analysis. Repeating our analyses described below grouping the descriptions by round rather than occurrence provides a pattern of results which are qualitatively similar.

2.3.1 Communicative success

This is simply the proportion of directed images which were successfully identified by the matcher(s). To count as a success in the Triad condition, both matchers needed to identify the correct image. As each matcher was free to interact with the director until they felt that they had identified the intended referent before ending each round, we expected that communicative success would be near ceiling from the outset as is typically the case in these paradigms (e.g., Clark and Wilkes-Gibbs 1986, where the matcher error rate was only 2%).

2.3.2 Description length

As is standard in the literature on the emergence of communicative conventions, we measured the length in characters of the descriptions produced, using the labels extracted as described above. This measured the efficiency of the developing communication systems.

2.3.3 Use of geometric descriptions

As discussed above, descriptions that make use of literal, as opposed to figurative, terms are likely to be more easily understood by a naive hearer (Fussell and Krauss 1989). We therefore assessed the use of geometric lexical items as a measure of description transparency—geometric terms are considered more literal as the stimuli were constructed from regular geometric shapes. A greater use of geometric shapes would indicate more transparent form-meaning mappings, considered an indication of lower linguistic complexity (Wray and Grace 2007; Lupyan and Dale 2010; Trudgill 2011).

For each description, we counted (automatically, using a search function) the number of times geometrical lexical items ('square', 'rectangle', 'triangle', 'diamond', 'trapezoid', and 'parallelogram') occurred. For instance, the description 'the camel with one hump' would have a geometric description score of 0, while 'dish from prev round, i think. diamond, then triangle attached to square on top of 3 overlapping triangles there is a similar one with dish unattached' would have a score of 4.[7]

2.3.4 Semantic specificity

We considered the minimum taxonomic depth of the description heads within the WordNet (WordNet 3.1 2010) hierarchy, to assess the claim that more esoteric communication (in this case that of the Dyad condition) would result in greater semantic complexity and more specific lexical items (Wray and Grace 2007; Trudgill 2011). As an example, the WordNet entry for 'animal' has depth 6: the shortest path of hyponyms from the entry at the top of the hierarchy has six steps (entity, physical entity, object, unit, living thing, organism, animal). The entry for 'pet' has depth 7, being a direct hyponym for 'animal'. We used this as a proxy for specificity, with 'pet' being a more specific term within a larger subset of 'animals'.

2.3.5 Description systematicity

Our set of twenty-four tangrams is organized into four subsets: animals, birds, people, and trinkets (Fig. 1). Our second measure of transparency attempted to capture whether this categorical structure in the set of referents was reflected in the set of descriptions the participants use to describe those referents; did participants use one term or a set of semantically related terms to describe all animal tangrams, a separate term or related set of terms for describing people, and so on? If so, the set of labels would systematically reflect the category structure in the underlying set of referents. Higher levels of systematicity may indicate more transparent form-meaning mappings, indicative of simpler language (Wray and Grace 2007; Lupyan and Dale 2010; Trudgill 2011).

In order to quantify the systematicity of sets of descriptions, we adapted the technique provided by Mantel (1967), which has been applied to measure systematic structure in artificial languages (e.g., Kirby et al. 2008). The intuition behind this measure is that, in a systematic language or set of descriptions, similar meanings (i.e., tangrams drawn from the same set) will be associated with similar descriptions (i.e., using terms with the same or similar semantics). We quantified this by evaluating the correlation between pair-wise differences in meaning and pair-wise differences in the associated descriptions—in a systematically structured set of descriptions, these two quantities would be correlated.

Quantifying systematic structure therefore required measures of difference between referents, and measures of distance between their descriptions. We used a simple measure of referent similarity: referents from the same (sub)set of tangrams were assigned a referent distance of 0, referents from different (sub)sets were assigned a referent distance of 1 (e.g., any two tangrams from the animal set had a difference score of 0, any animal had a difference score of 1 from any tangram from the person set). Our measure of distance in the descriptions produced by our participants was somewhat more complex, since we wanted to test for conceptual similarity in the description scheme mapping on to the categorical structure in the referent space, rather than strict string similarity as is often used in artificial language learning experiments. In order to quantify the conceptual distance between two descriptions, we therefore took their head words (as described above).

Each unique head (a total of 163 unique heads in a list of 1,330 heads overall) was checked against its WordNet entry.[8] The semantic distance between a pair of heads was calculated using path similarity: the shortest possible hypernym and hyponym path between two WordNet entries. This was scaled so that the maximum similarity between two entries was 1 (i.e., an entry is compared with itself), and the minimum was 0 (i.e., the two entries could not be further apart).[9] Conceptual distance between description heads was taken as 1 minus path similarity. Where path similarity was undefined, as was the case for pairs of particularly unrelated heads, such as 'silhouette' and 'blue', conceptual distance was taken as 1.

In order to measure the systematicity of a set of descriptions produced by a group, we calculated the distances between all pairs of tangrams and their associated descriptions, then took the Pearson's correlation between these two sets of distances. High r-values here were suggestive of systematicity, that is, referents from

the same category being described with conceptually similar descriptions. In order to evaluate the statistical significance of these *r*-values (calculated from non-independent sets of distance scores), we used a Monte Carlo simulation technique: we generated 10,000 randomized assignments of labels to stimuli (by simply shuffling the descriptions associated with the tangrams), and calculated *r* for each of those randomizations, giving us a distribution of *r* scores which would be expected for systems lacking systematicity (as was the case for our randomizations). We then calculated the *z*-score of the actual *r*-value: *z* greater than 1.96 indicated a degree of systematicity unlikely ($P < 0.05$) to arise in a non-systematic set of descriptions. Note that scores greater than 1.96 also suggested that our participants were sensitive to the category structure which we built into our set of tangrams; if participants were not sensitive to the categories, then the systematicity scores based on our groupings would have been random and so produce low structure scores.

2.4 Statistical tests

We performed a linear mixed effects analyses using R (R Core Team 2013) and *lme4* (Bates et al. 2013). Appropriate transformations and link functions were determined by visual inspection of the data for each analysis, and residuals were visually inspected for homoscedasticity. For the communicative success measure based on binomial data, we used logit regression; for the description length measure based on negatively skewed data, we used linear regression after log-transforming the data; for the use of geometric terms measure based on zero-inflated count data, we used Poisson regression; otherwise we used linear regression. As fixed effects, all analyses included Condition (Dyad or Triad, Dyad as intercept), Round, or Occurrence (−1, so that the intercept of the model represents Round 1 or Occurrence 1). The analysis of communicative accuracy included by-Group random intercepts and random slopes for Round; for the other measures, we included by-Group and by-Image random intercepts and random slopes for Occurrence for each.[10] In the linear regression models, we used *P*-values estimated from the resultant *t*-statistics, taking an upper bound for the degrees of freedom as the number of observations minus the number of fixed parameters in the model (Baayen 2008). For all analyses, we consider *P*-values < 0.05 as statistically significant.

2.5 Results

Average communicative success, length of description, geometric description score, semantic specificity (head WordNet depth), and semantic structure are illustrated in Fig. 4 for each condition.[11]

2.5.1 Communicative success

As expected (and intended as part of the experimental design), communicative success was near-ceiling throughout the experiment, and exhibited a very small increase over rounds, with 96% of directed images correctly matched in Round 1 rising to 100% in Round 6. We fit a logit linear regression to the communicative success data, as explained above: the full model was no better than the equivalent null model ($\chi^2(3) = 6.474$, $P = 0.091$), indicating that both dyads and triads were essentially at ceiling accuracy throughout.

2.5.2 Description length

We fit a linear model to the log-transformed description length data, which was significantly better than the null model ($\chi^2(3) = 40.26$, $P < 0.001$). There were significant effects of condition ($\beta = 0.456$, SE = 0.179, $t(880) = 2.54$, $P = 0.011$), and occurrence ($b = -0.375$, SE = 0.067, $t(880) = -5.60$, $P < 0.001$), and a (marginally) non-significant effect of their interaction ($b = -0.183$, SE = 0.094, $t(880) = -1.93$, $P = 0.054$). While descriptions in triads are generally longer, they do not remain longer than those of dyads; by Occurrence 4 there is no difference in mean description length ($t(9) = -0.154$, $P = 0.881$).

2.5.3 Use of geometric descriptions

In Occurrence 1, the average geometric description score was 0.833 (i.e., on average, most descriptions used a geometric term) in the Dyad condition and 1.392 in the Triad condition. These scores fell to 0.262 and 0.414, respectively, by Occurrence 4. The Poisson regression model was significantly better than the null model ($\chi^2(3) = 27.096$, $P < 0.001$), and there were significant effects of condition ($b = 0.456$, SE = 0.200, $z = 2.282$, $P = 0.022$) and occurrence ($b = -0.484$, SE = 0.129, $z = -3.738$, $P < 0.001$), but no effect of their interaction ($b = -0.154$, SE = 0.167, $z = -0.921$, $P = 0.357$). Triads used more geometric descriptions initially, use of geometric descriptions decreased over time in both conditions. There was no difference between conditions in the proportion of geometric descriptions used at Occurrence 4 ($t(9) = -1.033$, $P = 0.329$).

This is consistent with the description lengths analysis in the previous section. Since triads produce longer descriptions overall, the greater frequency of geometric terms in their descriptions may simply be a consequence of this greater length. Including

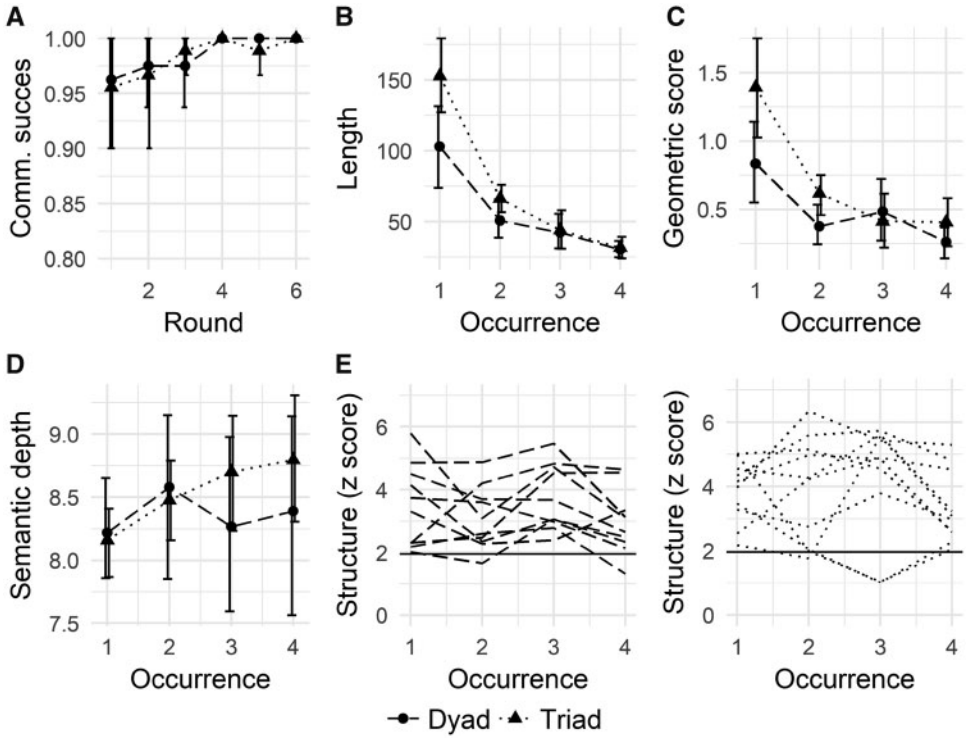

Figure 4. Experiment 1: (**A**) Communicative success by condition and round; (**B**) average length of descriptions, (**C**) geometric description scores, (**D**) semantic specificity, and (**E**) semantic structure by condition and occurrence. As intended, communicative success is near-ceiling throughout the experiment in both conditions. Descriptions shorten with occurrence in both conditions; triads initially produce longer descriptions, but these shorten more rapidly to produce descriptions of equivalent length across the two conditions by Occurrence 4. Use of geometric descriptions also decreases with occurrence in both conditions; triads initially use a greater number of geometric descriptions, but their use decreases to produce equivalent levels across the two conditions by Occurrence 4. There is no evidence of an effect of condition or occurrence on semantic specificity, and so no evidence of an effect on semantic specificity. For semantic structure, the horizontal line marks the critical *z*-score of 1.96. The heads are generally structured relative to the set of images throughout, but there is no effect of condition or occurrence. Error bars are 95% confidence intervals.

description length as a random intercept and occurrence as a by-description length random slope in the Poisson regression model described above resulted in a model significantly different from its null equivalent ($\chi^2(3) = 9.369$, $P = 0.025$), but only a better fit of the data under AIC (1,664 compared with 1,667) and not BIC (1,726 compared with 1,715).[12] In any case, under the full model there was no effect of condition ($b = 0.245$, SE $= 0.147$, $z = 1.670$, $P = 0.095$), occurrence ($b = -0.175$, SE $= 0.097$, $z = -1.794$, $P = 0.073$), or their interaction ($b = -0.047$, SE $= 0.107$, $z = -0.439$, $P = 0.660$). This suggests that the longer descriptions and greater use of geometric descriptions in the Triad condition are related; specifically, the slightly higher use of geometric terms will likely have led to longer descriptions, and once description length is controlled for, the difference in use of geometric terms disappears.

2.5.4 Semantic specificity

The full linear mixed model for average head depth was no better than its null model ($\chi^2(3) = 5.395$, $P = 0.145$). There is therefore no evidence to suggest that the descriptions in one condition are more semantically complex or specific than the other, nor indeed that semantic specificity changes over time.[13]

2.5.5 Description systematicity

Structure *z*-scores by occurrence are also illustrated in Fig. 4—recall that these reflect the extent to which the categorical structure of the tangrams are reflected in the heads of the participants' descriptions. As an example of a structured set of descriptions, one of the Dyad group's Occurrence 1 heads which referred to the animal images were 'emu', 'camel', and 'fox'; their bird descriptions were always headed by 'bird', their person description heads

were always 'person', and their trinket image heads were 'candle' (twice) or 'triangle'. This description scheme receives a high structure score because description heads are highly consistent within categories; even in the case of animal description heads, 'emu', 'camel', and 'fox' are semantically similar, and distinct from the descriptions for the other tangram categories. An example of an unstructured description set, which occurred in the Triad condition in Occurrence 3, is: 'camel', 'gesture', and 'throne' for the animals; 'eagle' (twice) and 'duck' for the birds; 'man' (twice) and 'chef' for the people; and 'man' and 'candle' for the trinkets. The lower structure score arises from the reduced consistency/similarity within each category, and some overlap between categories ('man' is used for both people and trinkets).

Strikingly, structure scores are high throughout, with all twenty groups obtaining systematicity scores reflecting a systematic, transparent mapping from referents to descriptions in Occurrence 1, and nineteen doing so in Occurrence 4. This indicates that our participants were generally sensitive to the category structure we built into the set of tangrams. The regression model on structure was not significantly better than the null model ($\chi^2(3) = 2.496$, $P = 0.476$). Hence, there is no evidence of a difference between conditions, or any difference in the systematicity of the description heads by occurrence.

2.6 Discussion

This study follows previous work in demonstrating that the communication of novel referents becomes more efficient with repeated use in dyads (Krauss and Weinheimer 1964; Clark and Wilkes-Gibbs 1986; Garrod et al. 2007), and shows the same behaviour in groups of three participants. Communicative accuracy remains high over repeated description of the tangrams, while the length of the descriptions reduces. Earlier descriptions in the Triad condition were longer than those in the Dyad condition; even the minimal increase in group size was enough to elicit a quantitative difference in the initial referring expressions. With repeated use, however, they became equally succinct.

We sought to test the hypothesis that group size (one of the features which distinguishes group types and communicative contexts on the esoteric/exoteric continuum) would influence the complexity of the emerging descriptive conventions. In this study, complexity would be evidenced by more compact descriptions, greater use of figurative rather than literal (geometric) descriptions, semantically more specific lexical items and less systematic referent-to-description mappings (Wray and Grace 2007; Trudgill 2011). There is no evidence in our data of an effect of group size on the final, Occurrence 4 description schemes arrived at in our groups. As discussed above, this could be due to the fact that our manipulation of group size is rather minimal compared with the range of social group sizes underpinning the esoteric/exoteric distinction in the wild. However, two of our measures do indicate effects of condition in the early stages of the negotiation process, where (in line with the predictions of these theories), triads use longer descriptions and make greater use of easy-to-identify geometric terms, which suggests that manipulations of this magnitude can influence the form of emerging communicative conventions, at least initially. There are no effects of condition or occurrence on systematicity or semantic specificity, suggesting that in our paradigm this is unaffected by group size, at least for the group size comparison we have considered here.

3. Experiment 2: the effect of shared knowledge

In Experiment 2, we adapted the methodology of Experiment 1 to test the claim that greater levels of communally shared knowledge can lead to more complex language, as argued by Wray and Grace (2007) and Trudgill (2011). In Experiment 1, all members of each group shared the same set of twelve non-target foil tangrams, which were possible selections by matchers but never the target of a director's description. In this experiment we manipulated the sharing of foils across participants while holding group size constant (looking only at triads): we compared the triads from Experiment 1 (providing our relatively esoteric baseline, which we will refer to here as the *Foils Shared* condition) with a new set of triads in which we reduce the amount of shared information by having foils unique to each member of the group (the *Foils Not Shared* condition). This comparison provided a test of the hypothesis that less communally shared information leads to lower complexity communicative conventions.

3.1 Materials and methods

3.1.1 Participants

Our participants in the Foils Shared condition were those detailed under Experiment 1, assigned to the Triad condition.

We ran an additional thirty-three participants (twenty-eight female, five male; aged between 18 and 40 years, mean 22.4) in the Foils Not Shared condition, again recruited via the Student and Graduate Employment Service at the University of Edinburgh.

These participants were paid £7 for around 60 min. Data from thirty participants (10 triads) were retained, the remaining data being discarded for failure to complete six rounds in the allotted time (three participants total, one triad).

3.1.2 Materials

The set of forty-eight tangrams used was identical to Experiment 1.

As in the Triad condition of Experiment 1, in the Foils Not Shared condition of Experiment 2, twelve tangrams were randomly selected for communication, nine of which were the target for description in any one round. In contrast to the Foils Shared condition, where twelve tangrams were selected as the foils for all three participants, the remaining thirty-six tangrams were equally and randomly divided between the participants to give each an idiosyncratic set of foils. Each individual participant's grid therefore contained twenty-four tangrams as before, but only the twelve selected for communication were the same across the three grids. The participants were not explicitly told that there were any differences between their sets of tangrams.

3.1.3 Procedure

The procedure was identical to the Triad condition of Experiment 1. We aimed to collect a minimum of six rounds of data, and groups who failed to reach this minimum were excluded from analysis.

3.2 Statistical tests

All coding and analysis was carried out as for Experiment 1, and we used the same five dependent variables. The Foils Shared condition was taken as the baseline in all analyses.

3.3 Results

The results are illustrated in Fig. 5.

3.3.1 Communicative success

Communicative success was again near ceiling, which was unsurprising given the participants' ability to continue interacting until the matchers believed they had accurately identified the directed images. Ninety-three percent of directed images were correctly matched in Round 1, rising to 97% in Round 6. The full model featuring condition and round was a significantly different fit of the data than the null model ($\chi^2(3) = 14.466$, $P = 0.002$). Under AIC, the model was a better fit of the data (316 compared with 325), but it was a worse fit under BIC (351 compared with 345). The full model indicated a significant effect of round ($b = 1.083$, SE $= 0.496$, $z = 2.186$, $P = 0.029$), but no significant effect of condition ($b = -0.932$, SE $= 0.750$, $z = -1.242$, $P = 0.214$) and no significant interaction between round and condition ($b = -0.766$, SE $= 0.435$, $z = -1.759$, $P = 0.079$): communicative success starts at similar levels and increases over rounds at similar rates in both conditions.

3.3.2 Description length

The full model fit for the log-transformed description length data was significantly better than the null model ($\chi^2(3) = 63.168$, $P < 0.001$). There were significant effects of condition ($b = 0.219$, SE $= 0.088$, $t(900) = 2.49$, $P = 0.013$) and occurrence ($b = -0.566$, SE $= 0.050$, $t(900) = -11.32$, $P < 0.001$), but no effect of the interaction between condition and occurrence ($b = -0.049$, SE $= 0.066$, $t(900) = -0.73$, $P = 0.466$). By Occurrence 4, there was no significant difference between the conditions ($t(9) = -0.603$, $P = 0.561$). Consistent with the results of Experiment 1, any differences between the conditions is eliminated by Occurrence 4.

3.3.3 Use of geometric descriptions

In Occurrence 1, the average number of geometric terms per description was 0.967 for the Foils Shared conditions and 1.144 for the Foils Not Shared condition. These scores fell to 0.333 and 0.189, respectively, in Occurrence 4. The full Poisson regression model was significantly better than the null model ($\chi^2(3) = 43.522$, $P < 0.001$). There was a significant effect of occurrence ($b = -0.608$, SE $= 0.091$, $z = -6.688$, $P < 0.001$), and a marginal effect of the interaction between condition and occurrence ($b = -0.211$, SE $= 0.108$, $z = -1.956$, $P = 0.051$), but no effect of condition ($b = 0.222$, SE $= 0.160$, $z = 1.390$, $P = 0.165$). Both conditions showed a decrease in number of geometric descriptions over round, but the comparatively exoteric Foils Not Shared condition lost geometric descriptions more rapidly than in the relatively esoteric Foils Shared condition; there was no significant difference between conditions in Occurrence 4 ($t(9) = 1.438$, $P = 0.184$).

The full Poisson regression model for geometric descriptions which also included description length as a random intercept and occurrence as a by-description length random slope was significantly better fit than its null equivalent ($\chi^2(3) = 41.558$, $P < 0.001$). The model indicated no effect of condition ($b = 0.254$, SE $= 0.164$, $z = 1.553$, $P = 0.120$), but there was an effect of occurrence ($b = -0.439$, SE $= 0.104$, $z = -4.226$, $P < 0.001$) and the interaction of condition and occurrence

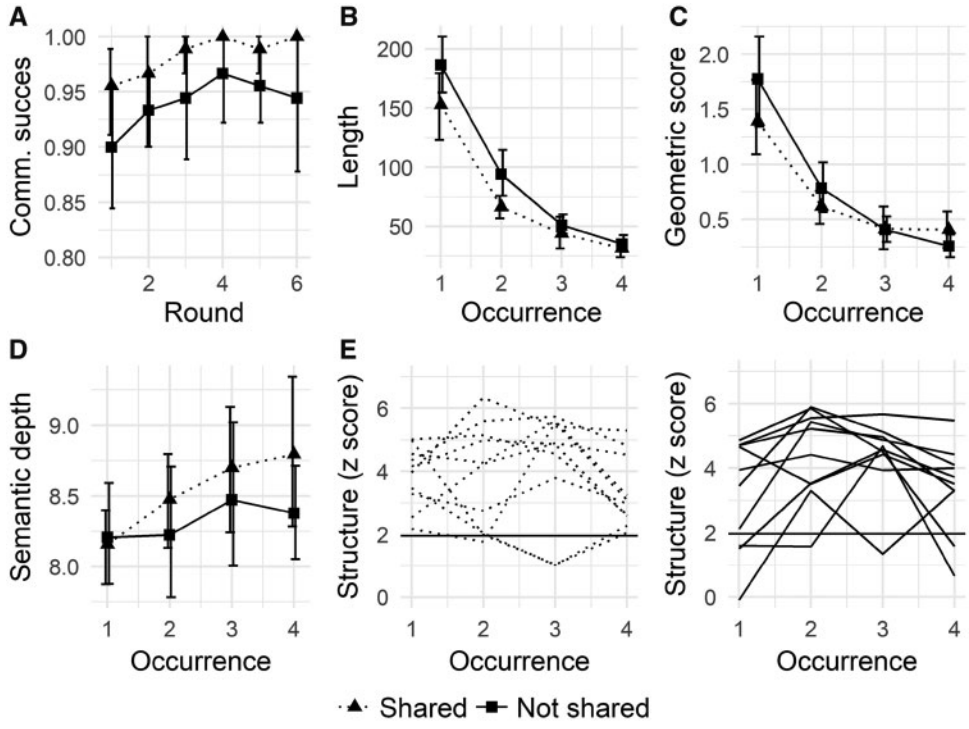

Figure 5. Experiment 2: (**A**) Communicative success by condition and round; (**B**) average length of descriptions, (**C**) geometric description score, (**D**) semantic specificity, and (**E**) semantic structure by condition and occurrence. As intended, communicative accuracy is high from Round 1, and increases as the experiment progresses. There is no difference between the conditions. Descriptions shorten with occurrence in both conditions; by Occurrence 4, the descriptions are of equivalent length across the two conditions. Use of geometric descriptions decreases with occurrence in both conditions, and to a greater extent in the Foils Not Shared condition; there is no difference between conditions by Occurrence 4. There is no evidence of an effect of condition on depth within the WordNet hierarchy, and so no evidence of an effect on semantic specificity. There is some limited evidence that semantic complexity increases with occurrence, however. For semantic structure, the horizontal line marks critical *z*-score. The heads are generally structured relative to the set of images throughout, but there is no effect of condition or occurrence. Error bars are 95% confidence intervals.

($b = -0.271$, SE $= 0.098$, $z = -2.763$, $P = 0.006$). Therefore, there is some indication that the Foils Not Shared descriptions shed geometric terms per character more rapidly, providing weak evidence that this difference may not purely derive from differences in overall length of referring expression.

3.3.4 Semantic specificity

There was only marginal evidence that the model for head specificity[14] was different to the null model ($\chi^2(3) = 7.416$, $P = 0.060$); under AIC a better fit of the data (though only by 1; 3,840 compared with 3,841), but under BIC a worse fit (3,892 compared with 3,880). The model indicated a significant effect of occurrence ($b = 0.206$, SE $= 0.083$, $t(889) = 2.486$, $P = 0.013$), but no effect of condition ($b = -0.039$, SE $= 0.304$, $t(889) = -0.128$, $P = 0.898$), or the interaction of occurrence and condition ($b = -0.110$, SE $= 0.118$, $t(889) = -0.931$, $P = 0.352$). As in Experiment 1, there was a lack of an effect of condition on semantic complexity, but here there was some limited evidence that semantic complexity increases with occurrence.

3.3.5 Description systematicity

As in both conditions in Experiment 1, the average *z*-scores in the new Foils Not Shared condition were consistently greater than 1.96, suggesting the sets of descriptions were significantly structured throughout, and that the participants were sensitive to the structure which we built into the tangram sets. The full model was not a significantly better fit to the data than the null model ($\chi^2(3) = 0.631$, $P = 0.889$), and so there was no evidence for an effect of condition or occurrence on systematicity.

3.4 Discussion

The results of this experiment largely mirror those of Experiment 1: again the comparatively exoteric

condition (here, the Foils Not Shared condition) results in longer descriptions in the earlier rounds, but with that difference being eliminated through repeated use. There is also again little evidence of any effects of exotericity on description transparency; geometric terms are more rapidly lost in the Foils Not Shared than in the less exoteric Foils Shared condition with no difference between conditions by Occurrence 4; similarly, while the sets of descriptions in both conditions are highly systematic, there is no evidence for any difference between conditions in the level of transparency or semantic specificity. In sum, there is therefore little evidence that our esotericity/exotericity manipulation in this experiment impacted on the complexity of language use, beyond the very early stages of the grounding process.

4. Experiment 3: transparency to naive observers

Finally, we ran an additional experiment in order to test the transparency of the descriptions of Experiments 1 and 2, by seeing how well naive raters could match descriptions to their referents (following, e.g., Fay et al. 2008). In removing the shared knowledge established through the grounding of the descriptions, we could more directly assess the claim that more exoteric communication leads to more transparent form-meaning mappings (Wray and Grace 2007). Under this hypothesis, we expected naive individuals to more accurately match the descriptions produced by Triad Foils Shared groups to their intended images, compared with the descriptions produced by Dyads. Similarly, descriptions produced in the Triad Foils Not Shared condition should have been more transparent and interpretable than those of the Triad Foils Shared.

4.1 Materials and methods

4.1.1 Participants

A total of 345 participants were recruited on CrowdFlower[15] and required to match descriptions to images, 330 rated 12 descriptions each, and 15 rated 6 descriptions each. We paid $0.20 for each participant's contribution.

4.1.2 Materials

We considered the Occurrence 4 descriptions across the three conditions of Experiments 1 and 2, with some minor alterations to the descriptions so as not to unnecessarily confuse the raters. All references to previous labelling of the image or use of the description were removed, including, for example, 'AGAIN', 'thing XXXX got confused with', 'from first round', 'we described that one as', and 'same'. References to participant names or usernames (already marked by a series of 'X's) were removed. Descriptions were de-pluralized where they had been used to refer to multiple images. Finally, three labels were excluded in case they caused offence (e.g., 'dinosaur with dick out'). This left a total of eighty-four descriptions from the Dyad condition, ninety-six from the Triad with Foils Shared, and ninety from the Triad with Foils Not Shared: 270 descriptions in total.

4.1.3 Procedure

The testing trials were randomly distributed across participants. For a given description, the participant was presented with an array of twenty-four images, the same seen by a matcher during the experiment. In the Dyadic condition, this was the other person. In the Triadic conditions, one of the two matcher arrays was randomly selected. The arrays were presented in the same order, but what would have been the director images in the experiment were not marked (i.e., this meant that the CrowdFlower participant could select any of twenty-four images, whereas the participants in Experiments 1 and 2 were not allowed to select the three or four images they were allocated to direct themselves).

4.2 Results

Average accuracy for a single description ranged from 0% to 93%, indicating that some descriptions were never matched to their intended image, while others were very accurately matched. Overall accuracy, the mean of the averages for each description, was 51%: 48% for the descriptions produced in the Dyad conditions, 54% for the Triad Foils Shared, and 49% for the Triad Foils Not Shared. Chance performance was 4%. Correct identification of individual tangrams ranged from 15% (for a Bird with four unique descriptions) to 82% (an Animal with seven unique descriptions). Where an Animal tangram was the intended referent, accuracy was 51% (the average of the average accuracy scores for all unique descriptions intended to describe Animals); Birds 49%; People 51%; Trinkets 44%.

A linear mixed model with logistic link was constructed with condition as a fixed effect, with by-Rater and by-Intended Image random intercepts and random slopes for condition for each. Condition was Helmert contrast coded, allowing two contrast types to be investigated: Triad Foils Shared (the baseline) versus Triad Foils Not Shared, followed by Triads versus Dyads. The

model was not better than the equivalent null model ($\chi^2(2) = 2.001$, $P = 0.368$). There is therefore no evidence that the descriptions were more accurately matched in either of the Triad Foils Shared or Triad Foils Not Shared conditions, or in the Triad conditions compared with the Dyads.

Experiment 3 therefore provides no support for the view that more exoteric communication results in more transparent form-to-meaning mappings (Wray and Grace 2007), whether exotericity is manipulated by the amount of shared knowledge shared by members of a group, or group size.

5. General discussion

Experiments 1 and 2 replicate the findings of previous studies (Krauss and Weinheimer 1964; Clark and Wilkes-Gibbs 1986) in that the length of referring expression decreases as participants repeatedly describe and match descriptions of stimuli through interaction. In the earlier interactions, we see longer descriptions in the conditions which have a feature typical of more exoteric communicative contexts: larger group size or lower levels of shared information. However, these differences between conditions disappear over repeated interaction. There is also no evidence in the final (Occurrence 4) descriptions of condition-dependent differences of semantic complexity, use of literal terms, or of transparency of form-meaning mappings between the descriptions and the semantic space. Experiments 1 and 2 therefore provide little evidence to support the view that more esoteric communicative contexts could lead to languages being more efficient, having less transparent form-meaning mappings, or using more highly specific lexical items (Wray and Grace 2007; Trudgill 2011).

Experiment 3 also provides no evidence that larger groups, or groups that have a greater amount of information shared between its members, may develop expressions which are more easily interpreted by individuals not party to the negotiation process, and so offers no support for the hypothesis that more exoteric communication may ease comprehension for out-group members (Wray and Grace 2007).

We cannot of course rule out that our experimental design here has failed to capture genuine effects of group size and shared knowledge. It is possible that our experiments suffer from a lack of power, and that we may have found differences between our conditions with larger sample sizes. The contrast between our conditions may also be too subtle; as noted earlier, in the real world the contrast between esoteric and exoteric contexts would be much larger. Our experiments also involve the communication of only a small set of referents, certainly compared with real-world human communication systems. If these experiments were repeated with much larger group size differences in Experiment 1, or if the ratio of foils to potential targets was much larger in Experiment 2, then the condition-dependent differences we see in the initial sets of descriptions may have more lasting effects. This might be more likely if the number of referents was substantially increased as well. We therefore suggest that 'scaling up' these experiments here may be worthwhile, particularly as Fay et al. (2008) have illustrated how greater transparency can persist in larger groups (albeit in groups where individuals interacted dyadically) in their graphical communication study contrasting dyads with groups of eight.

It is also worth noting some other limitations of these experiments relative to the literature discussed in Section 1. We have only manipulated two of the factors which Wray and Grace (2007) and Trudgill (2011) suggest characterize esoteric and exoteric groups and communicative contexts, and we have only done so considering stable, closed groups of interacting participants. Future experimental work could include the manipulation of characteristics of human social groups other than simply their size (such as the strength of the social connections between the individuals in a group; see, e.g., Milroy 1980, for discussion of the effect of different social structures on language change). Manipulating multiple factors which are characteristic of more or less esoteric groups would also be a worthwhile avenue of research, particularly if, as argued by Trudgill (2011), group type effects on language features may be driven by the *interaction* of different social factors. It may be that group size by itself, as we manipulated in Experiment 1, is not enough to distinguish comparative esoteric and exoteric communicative contexts; larger group size may primarily be relevant in increasing the amount of (exoteric) communication between strangers.

Future experiments could also consider alternative interpretations of how different groups could have different degrees of 'shared knowledge'. For example, a similar experiment to those presented here could compare the descriptions of groups in which the participants knew each other well (a 'society of intimates') with those of complete strangers. We also stress that though we have investigated the effect of two factors which contribute to a communicative context being more or less esoteric and measured whether this affects description length, the transparency of form-meaning mappings, and the use of more semantically complex lexical items, all of which have been argued to contribute to language complexity, we have by no means exhausted all complexity relevant language features here. Experimentally investigating the

effects of group size and network structure, and different types of shared information, on features such as morphological complexity, would be particularly worthwhile. We suggest that these features may be better investigated using an artificial language learning paradigm, however, rather than the natural language referential communication designs we have used here.

Ultimately, there is no evidence here that the processes of grounding between group members may be a mechanism by which esoteric communication could lead to lower levels of transparency, and hence greater linguistic complexity. Instead, as argued in Atkinson et al. (2018), if interaction between speakers does systematically influence linguistic complexity, it may be in spreading existing simplifications which arise as a result of adult learning.

6. Conclusion

We manipulated two different social factors and investigated how each influenced language complexity. The manipulations of group size and amount of communally shared information in Experiments 1 and 2 show no evidence of lasting effects of esotericity on language complexity: while more exoteric communicative contexts initially lead to longer descriptions and greater use of more literal descriptive terms, this effect is eliminated with repeated interaction. Experiment 3 then finds no effect of either manipulation on the interpretability of the emergent conventions by out-group members, and so no evidence that the communicative pressures of more exoteric social groups may lead to more transparent lexical items.

Acknowledgements

We thank three anonymous reviewers for their helpful comments on an earlier draft of this manuscript.

Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 681942, and a Marie Curie IOF (PIOF-GA-2009-236632-ERIS). The first author was supported by an Arts & Humanities Research Council PhD Studentship (AH/K503010/1).

Conflict of interest statement. None declared.

Notes

1. The complete set of tangrams is available at <http://dx.doi.org/10.7488/ds/1979> accessed 25 Sep 2018.
2. Available at <http://cogsci.eecs.qmul.ac.uk/diet/> accessed 25 Sep 2018.
3. This surrounding material included turn negotiation (e.g., 'Shall we complete one person's set first?'), and text not directly related to the task of directing and matching the images, such as responses to the round scores (e.g., 'hashtag amazing').
4. Occasionally, a trimmed description referred to two images which could not be separated, for example, 'both of the giraffes'. In such cases, the description line was considered (part of) the description for each image, but with lexical markers of plurality removed.
5. The complete set of descriptions, along with the description heads, lengths, head depths, and number of geometric terms (described below), is available at <http://dx.doi.org/10.7488/ds/1979> accessed 25 Sep 2018.
6. Extraction of the descriptions and heads from the surrounding material was done by the first author, who was not blind to experimental condition. However, this extraction process is essentially mechanical and seldom involved subjective judgements, and therefore we did not do additional blind coding.
7. As one reviewer pointed out, it may also make sense to include some lexical items which describe the relationship between the geometric shapes, such as 'overlapping', 'attached', and 'unattached', here. The inclusion of these terms makes no difference to the pattern of results we present below, however.
8. Where more than one entry existed, the most appropriate was identified. Two entries ('batman' and 'birdview'), the heads for a total of four descriptions, had no appropriate WordNet entry, and so these were removed from the analysis (0.3% of the data). Eight heads ('abstract', 'blue', 'fishy', 'hard', 'last', 'similar', 'upright', and 'wrong'), accounting for twelve descriptions in total, were coded as adjectives and a depth value of 0 was returned from WordNet in each case; these were removed from the analysis (0.6% of the data).
9. Python implementation details available at <http://www.nltk.org/howto/wordnet.html> accessed 25 Sep 2018.
10. Although some participants were recruited individually and others in groups of two or three, as described in Section 2.1, a self-selection variable was not included in our models. Although self-selected groups would have some shared communication history, which arguably could have had some influence on participant behaviour in the task, we could not quantify that shared history in a satisfactory way and thus did not include it in the analyses.

11. Detailed summaries of our analyses discussed below (along with those for Experiments 2 and 3) are available at <http://dx.doi.org/10.7488/ds/1979> accessed 25 Sep 2018.
12. Note that BIC will penalize additional model parameters to a greater extent. Under AIC, the penalty for k additional parameters is $2k$; under BIC, it is $\ln(n)k$, where n is the number of data points.
13. It is possible that our use of WordNet depth as a proxy for semantic specificity may have been too crude and that it does not accurately represent human judgements of specificity (see Wang and Hirst 2011). As a check, we gave a randomly shuffled list of the 163 unique heads to three naive raters, and asked them to rate each item for specificity on a seven-point scale. Taking an average of the three ratings for each head, we created a set of judgement ratings. These judgements at least correlated with the WordNet depths ($r = 0.54$, $P < 0.001$), and so we have no reason to suppose that use of WordNet depths was an inappropriate measure of semantic specificity here.
14. Four heads ('batman', 'firepit', 'he', and 'toblerone'), the heads for a total of twelve descriptions, had no appropriate WordNet entry, and so these were removed from the analysis. Three heads ('last', 'upright', and 'wrong'), accounting for three descriptions, were coded as adjectives and a depth value of 0 was returned from WordNet. These were removed from the analysis.
15. <http://www.crowdflower.com/> accessed 25 Sep 2018.

References

Atkinson, M., Kirby, S., and Smith, K. (2015) 'Speaker Input Variability Does Not Explain Why Larger Populations Have Simpler Languages', *PLoS One*, 10/6: e0129463.

——, Smith, K., and Kirby, S. (2018) 'Adult learning and language simplification', *Cognitive Science*. Doi: 10.1111/cogs.12686.

Baayen, H. (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Bates, D., Maechler, M., and Bolker, B. (2013) *lme4: Linear Mixed-Effects Models using S4 Classes*. Retrieved from <http://cran.r-project.org/package=lme4> accessed 25 Sep 2018.

Beckner, C. et al. (2009) 'Language Is a Complex Adaptive System: Position Paper', *Language Learning*, 59(Suppl. 1): 1–26.

Bentz, C., and Winter, B. (2013) 'Languages with More Second Language Learners Tend to Lose Nominal Case', *Language Dynamics and Change*, 3: 1–27.

Branigan, H. P., Catchpole, C. M., and Pickering, M. J. (2011) 'What Makes Dialogues Easy to Understand?', *Language and Cognitive Processes*, 26/10: 1667–86.

Caldwell, C. A., and Smith, K. (2012) 'Cultural Evolution and Perpetuation of Arbitrary Communicative Conventions in Experimental Microsocieties', *PLoS One*, 7/8: e43807.

Christiansen, M. H., and Chater, N. (2008) 'Language as Shaped by the Brain', *Behavioral and Brain Sciences*, 31/5: 489–508 (discussion 509–58).

Clahsen, H. et al. (2010) 'Morphological Structure in Native and Nonnative Language Processing', *Language Learning*, 60: 21–43.

Clark, H. H., and Wilkes-Gibbs, D. (1986) 'Referring as a Collaborative Process', *Cognition*, 22: 1–39.

Croft, W. (1995) 'Autonomy and Functionalist Linguistics', *Linguistic Soceity of America*, 71/3: 490–532.

—— (2000) *Explaining Language Change: An Evolutionary Approach*. London: Longman.

Dale, R., and Lupyan, G. (2012) 'Understanding the Origins of Morphological Diversity: The Linguistic Niche Hypothesis', *Advances in Complex Systems*, 15: 1150017.

Fay, N., Garrod, S., and Roberts, L. (2008) 'The Fitness and Functionality of Culturally Evolved Communication Systems', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 363/1509: 3553–61.

—— et al. (2010) 'The Interactive Evolution of Human Communication Systems', *Cognitive Science*, 34/3: 351–86.

Fox Tree, J. E. (1999) 'Listening in on Monologues and Dialogues', *Discourse Processes*, 27: 35–53.

——, and Mayer, S. A. (2008) 'Overhearing Single and Multiple Perspectives', *Discourse Processes*, 45/2: 160–79.

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*, 25/3: 203–19.

Galantucci, B., and Roberts, G. (2012) 'Experimental Semiotics: An Engine of Discovery for Understanding Human Communication', *Advances in Complex Systems*, 15: 1150026.

Garrod, S. et al. (2007) 'Foundations of Representation: Where Might Graphical Symbol Systems Come from?', *Cognitive Science*, 31/6: 961–87.

Givón, T. (1979) *On Understanding Grammar*. New York (NY): Academic Press.

Hupet, M., and Chantraine, Y. (1992) 'Changes in Repeated References: Collaboration of Repetition Effects?', *Journal of Psycholinguistic Research*, 21/6: 485–96.

Kirby, S. (1999) *Function Selection and Innateness: The Emergence of Language Universals*. Oxford: Oxford University Press.

——, Cornish, H., and Smith, K. (2008) 'Cumulative Cultural Evolution in the Laboratory: An Experimental Approach to the Origins of Structure in Human Language', *Proceedings of the National Academy of Sciences of the United States of America*, 105/31: 10681–6.

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: 113–4.

——, and —— (1966) 'Concurrent Feedback, Confirmation, and the Encoding of Referents in Verbal Communication', *Journal of Personality and Social Psychology*, 4/3: 343–6.

Lupyan, G., and Dale, R. (2010) 'Language Structure Is Partly Determined by Social Structure', *PLoS One*, 5/1: e8559.

——, and —— (2016) 'Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity', *Trends in Cognitive Sciences*, 20/9: 649–60.

Mantel, N. (1967) 'The Detection of Disease Clustering and a Generalized Regression Approach', *Cancer Research*, 2: 209–20.

Milroy, J. (1980) *Language and Social Networks*. Oxford: Blackwell Publishing.

Nettle, D. (1999) 'Is the Rate of Linguistic Change Constant?', *Lingua*, 108/2–3: 119–36.

—— (2012) 'Social Scale and Structural Complexity in Human Languages', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 367/1597: 1829–36.

Nichols, J. (2009) 'Linguistic Complexity: A Comprehensive Definition and Survey', in Sampson G., Gil D., and Trudgill P. (eds) *Language Complexity as an Evolving Variable*, pp. 110–25. Oxford: Oxford University Press.

R Core Team. (2013) *R: A Language and Environment for Statistical Computing*. Retrieved from <http://www.r-project.org/> accessed 25 Sep 2018.

Rogers, S. L., Fay, N., and Maybery, M. (2013) 'Audience Design through Social Interaction during Group Discussion', *PLoS One*, 8/2: 1–7.

Sapir, E. (1912) 'Language and Environment', *American Anthropologist*, 14/2: 226–42.

Schober, M. F., and Clark, H. H. (1989) 'Understanding by Addressees and Overhearers', *Cognitive Psychology*, 21/2: 211–32.

Sinnemäki, K. (2009) 'Complexity in Core Argument Marking and Population Size', in Sampson G., Gil D., and Trudgill P. (eds) *Language Complexity as an Evolving Variable*, pp. 126–40. Oxford: Oxford University Press.

Smith, K., and Kirby, S. (2008) 'Cultural Evolution: Implications for Understanding the Human Language Faculty and Its Evolution', *Philosophical Transactions of the Royal Society B, Biological Sciences*, 363/1509: 3591–603.

Trudgill, P. (2011) *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Wang, T., and Hirst, G. (2011) 'Refining the Notions of Depth and Density in WordNet-based Semantic Similarity Measures', *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1003–1011.

WordNet 3.1. (2010) 'About WordNet', in *WordNet*. Princeton University. Retrieved from <https://wordnet.princeton.edu/citing-wordnet> accessed 25 Sep 2018.

Wray, A., and Grace, G. W. (2007) 'The Consequences of Talking to Strangers: Evolutionary Corollaries of Socio-cultural Influences on Linguistic Form', *Lingua*, 117/3: 543–78.

Yoon, S. O., and Brown-Schmidt, S. (2014) 'Adjusting Conceptual Pacts in Three-Party Conversation', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40/4: 919–37.