Psychology of Language and Communication 2016, Vol. 20, No. 2

DE GRUYTER OPEN

DOI: 10.1515/plc-2016-0008

GARETH ROBERTS [1], BRUNO GALANTUCCI [2, 3]
[1] University of Pennsylvania, Philadelphia
[2] Yeshiva University, New York
[3] Haskins Laboratories, New Haven, Connecticut

INVESTIGATING MEANING IN EXPERIMENTAL SEMIOTICS

Experimental semiotics is a new discipline developed over the last decade to study human communication. Studies within this discipline typically involve people creating novel signs by associating signals with meanings. Here we suggest ways this discipline can be used to shed light on how people create and communicate meaning. First we present observations drawn from studies in which participants not only construct novel signals, but also have considerable freedom over what these signals refer to. These studies offer intriguing insight on non-saussurian signs (where a single unit of meaning is associated with different signals), communicative egocentricity, private and public meaning, and the distinction between meaningful and meaningless units in linguistic structure, that is between morphemes and phonemes (or analogous entities). We then present a novel quantitative approach to determining the extent to which a signal unit is meaningful, and illustrate its use with data from a study in which participants construct signals to refer to predetermined meanings. Aside from these specific contributions, we show more generally how challenging investigating meaning in Experimental Semiotics is, but we argue that this reflects the difficulties we must face when studying meaning, outside the lab as well as in it.

Key words: experimental semiotics, origins of meaning, semantics, pragmatics, language emergence

Over the last decade a number of researchers have developed a new approach to studying human communication (for a review, see Galantucci, Garrod, & Roberts, 2012). This approach, which has been labeled Experimental Semiotics (Galantucci, 2009; henceforth *ES*), offers researchers the opportunity to study novel communication systems created by participants under controlled conditions (henceforth *laboratory languages*). In what follows we will briefly illustrate

Address for correspondence: Bruno Galantucci, Department of Psychology, Yeshiva University, 2495 Amsterdam Avenue, New York, NY, 10033, USA. E-mail: bruno.galantucci@yu.edu

how this approach works and then provide some preliminary evidence suggesting that it can offer valuable insights into meaning in human communication This paper is divided into three sections. In the first section, we introduce the three main categories of studies performed by experimental semioticians, briefly sketching how they differ with respect to constraints on meaning. In Section 2 we focus on one of the three categories—the one that least constrains meaning—and present observations drawn from studies in this category concerning the emergence of meaning in human communication. In Section 3 we present analyses of data collected in a different category—in which meaning is more constrained—and argue that they corroborate the observations made in Section 2.

1. Categories of ES study

In any ES study participants engage in a task that involves creating *signs* by associating some kind of communicative behavior (such as drawing an image, or typing letters; henceforth *signal*) with a meaning (henceforth *referent*).[1] Beyond this, however, there are differences between studies in the extent to which participants are made explicitly aware of this task and how they might accomplish it. These differences have important implications for how easy it is to analyze meaning in the resulting data. Galantucci et al. (2012) distinguished three categories of ES study: *semiotic matching games, semiotic coordination games*, and *semiotic referential games*. In semiotic matching games, participants are provided by the experimenters with both referents and signals. On the face of it this might be expected to considerably simplify the analysis of meaning in such studies. However, as Galantucci et al. (2012, 480) noted, this is in fact a very heterogeneous category in which such simplification is introduced to allow other goals and dynamics to be explored, complicating the analysis. In Roberts's (2010) study of new-dialect formation, for instance, participants used an artificial language to negotiate with each other for resources, while Kirby, Cornish, and Smith's (2008) study of emergent structure did not even involve direct communication between participants. Because of such complications the remainder of this paper will focus on the other two categories.

In semiotic coordination games the participants' primary goal is to coordinate with other participants. Although this goal can be achieved only through communication, it is not always obvious to participants that such a route is available, and in some cases it might even be unclear to what extent participants' behavior is intended to be meaningful. This issue was particularly salient in an experiment by Scott-Phillips, Kirby, and Ritchie (2009), who explicitly studied the transition from non-communicative to communicative behavior and therefore

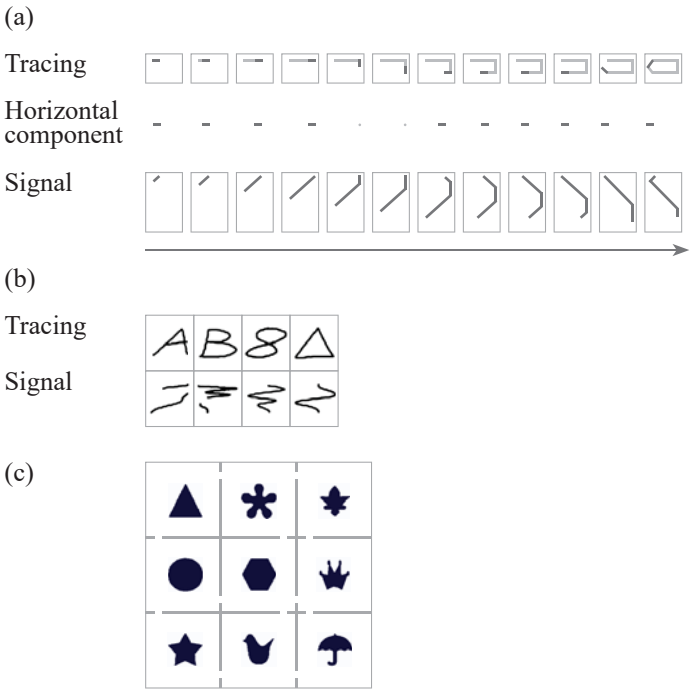
[1] One or two interesting studies are exceptions to this, since they employ ES methodology, but do not include referents at all and instead investigate the cultural evolution of meaningless communication-like behavior (Cornish, Christiansen, & Kirby 2010; Verhoef 2012).

did not provide a dedicated communication channel at all—unsurprisingly, participants found the task very challenging. Other semiotic coordination games have involved a distinct communication channel, although even in these studies participants find the task of establishing a new communication system to be a challenging one (e.g., Galantucci, 2005). Because communication in this category is secondary (if essential) to the primary goal of coordination, participants have a high degree of freedom not only in developing signals, but also in choosing what to refer to. Such studies therefore have the potential to offer a number of insights into how meaning emerges in novel communication systems. Several such insights will be discussed in Section 2. In spite of this potential, however, the freedom allowed participants in semiotic coordination games means that this is not necessarily an ideal paradigm for studying meaning in a controlled fashion. A much more controlled paradigm is provided by semiotic referential games, in which participants are provided with a closed set of referents and their primary goal is explicitly to develop signals to communicate the referents in question. Pictionary games (see, e.g., Garrod, Fay, Lee, Oberlander, & MacLeod, 2007; Theisen, Oberlander, & Kirby, 2010) are an obvious example. Studies in this category have already shed interesting light on meaning. Garrod et al. (2007), for instance, investigated the conditions under which iconic signs become symbolic, while Theisen et al. (2010) investigated the emergence of morphological structure in sign systems. In Section 3 we will use quantitative methods to analyze data from a semiotic referential game carried out by Roberts and Galantucci (2012), supporting the observations made in Section 2.

2. Insights from semiotic coordination games

Semiotic coordination games, as developed by Galantucci and colleagues (Galantucci, Fowler, and Richardson, 2003; Galantucci, 2005), typically work as follows. Pairs of participants (henceforth *dyads*) play a cooperative coordination game with interconnected computers. The game requires players to communicate in order to find one another in a virtual maze, but prevents the use of pre-established means of communication such as speech or writing. Players play the game over the internet (from different physical locations) and can communicate only by making tracings on a digitizing pad that are systematically transformed, in real time, into on-screen signals. This transformation is designed to prevent the use of standard graphic forms such as letters or numerals (see Figure 1a–b and Section 3.1 for a more detailed description), meaning that players must craft a novel visual communication system from scratch. In particular, to find one another in an efficient manner, they must craft a system capable of distinguishing the different locations in the maze (Figure 1c). Most dyads find the task challenging, but succeed at the game, developing relatively elaborate sign systems (Galantucci, 2005).

Figure 1. (a) How the tracings players produced on the digitizing pad appeared on screen; (b) how common graphic symbols drawn on the digitizing pad appeared on the screen (adapted from Galantucci & Garrod, 2010); (c) locations in one of Galantucci's (2005) mazes



2.1. Observations relevant to the study of meaning

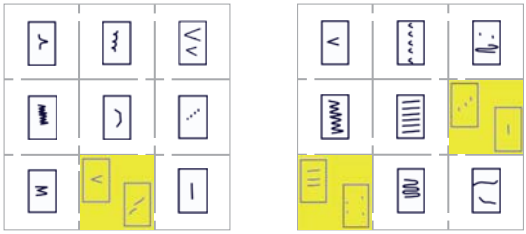
Several of the communication systems developed in semiotic coordination games have features that are particularly interesting for the study of meaning. These features will be described in Sections 2.1.1 to 2.1.4.

2.1.1. Non-saussurean signs

In some of the dyads studied by Galantucci (2005), the two players developed systems which included different signals for the same referent. Figure 2 presents two such systems developed for nine-location mazes, with the signals arranged in correspondence to the locations they refer to. As illustrated in the figure, for most of the locations the two players used the same signal. However, in some cases (highlighted in yellow in the figure) they used different signals. Such signals—which have been referred to in the literature as non-saussurean signs (Hurford, 1989)—were understood by the player who did not produce them,

leading to successful acts of communication. Yet they remained distinct for the entire time that players were in the nine-location maze. This is an interesting phenomenon. In natural language, speakers do sometimes communicate with each other using different words for the same thing, but this is usually a matter of speaking different regional varieties (e.g., an American talking to a British person might refer to a "sidewalk", which would be understood by the other as referring to a "pavement"). Examples not due to regional variation tend to be associated with such social factors as class, age, and gender, and are really another case of contact between different varieties. It is rather more peculiar to find cases not due to social or geographical factors, but they have been observed in relatively new sign languages such as Al-Sayyid Bedouin Sign Language (ABSL) (Sandler, Aronoff, Meir, & Padden, 2011), suggesting that non-saussurean signs might be part of the natural processes that lead to the creation of full-blown languages.

Figure 2. Signals developed by two dyads to refer to a maze with nine locations. For most locations, the signals are the same for the two players (i.e., there is only one signal per location) but for some locations (highlighted in yellow) they are not (i.e., one player uses one signal, while the partner uses another)



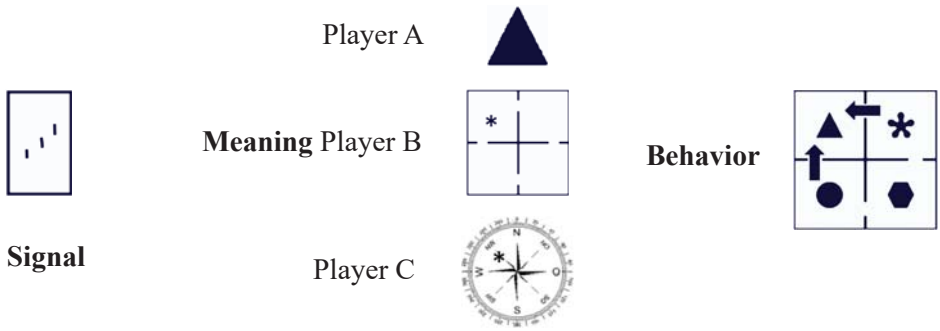
2.1.2. Saussurean signs with an egocentric component

Some players developed signals which looked the same for the two players on the screen but had slightly different meanings for the two players because of private information known only to one of them. For example, a player indicated the direction of the agent's vertical movements in the maze by drawing a vertical line from the bottom to the top of the pad for upward movements, and from the top to the bottom for downward movements. Given the properties of the communication device, the two lines appeared identical on the screen. Yet the player kept using them to differentially indicate direction, and was frustrated by her partner's "lack of understanding". Given that she was aware of her own intentions in creating the signals, the player expected the partner to understand them accordingly. In fact, her partner understood all that there was to understand in such conditions. To her, the signal meant vertical motion, with no indication of direction.

2.1.3. Difference between private and public meaning

Some players developed signals that elicited the same game moves (we refer to these as having the same *public meaning*) but meant different things to each of the players (we refer to these signals as having different *private meanings*). These differences—which were discovered only through interviews with the players—could be stark and widespread. For example, in a version of the game played by triads in a 22 maze (Galantucci & Roberts, 2012), the players in one of the triads developed a signal which had the same public meaning but three different private meanings. The signal, composed of three vertical dashes, is illustrated in Figure 3. To the first player the signal meant "triangle" (an icon which marked the maze location referred to by the signal), to the second it meant "the top left corner of the maze", and to the third it meant "north-west". This widespread misalignment in private meaning did not interfere with performance in the game: When the players saw the signals composed of three dashes, they all interpreted it as referring to the same maze location.

Figure 3. One signal with different meanings for different players

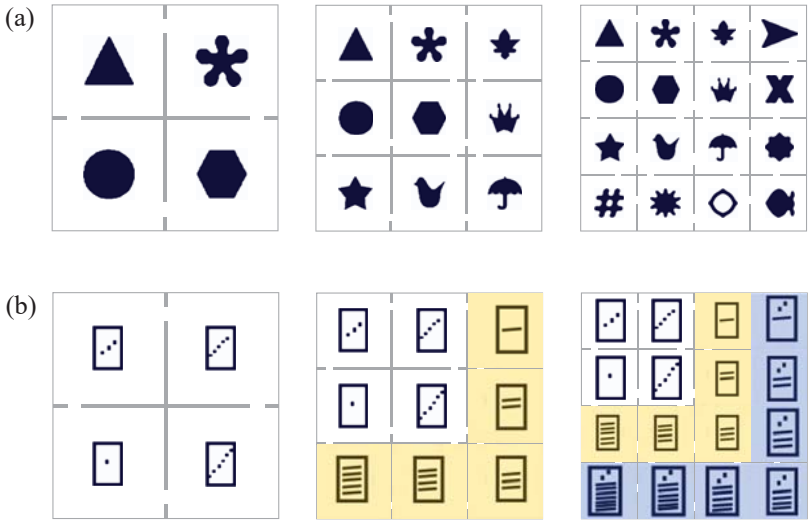


2.1.4. Blurred distinction between meaningful and meaningless units

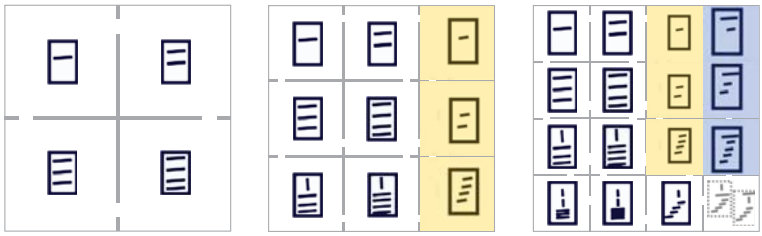
Another observation from Galantucci's (2005) study was that, in an analysis of the internal structure of signals, the distinction between meaningless units (the equivalent of phonemes in spoken languages) and meaningful units (the equivalent of morphemes in spoken languages) is often blurred. This phenomenon is illustrated in Figure 4. The figure presents two sign systems which were developed over time by two dyads playing three successive stages of the game, over which the maze grew in the manner depicted in Figure 4a.

The first sign system (Figure 4b) began at the first stage of the game with signals composed of short vertical dashes indicating the number of vertices of the icon which marked the maze (e.g., three dashes for a triangle). At this stage, the dashes referred to vertices and could be considered primordial forms of morphemes combined via a simple repetition strategy. At the second stage, the new locations in the game were indicated by signals composed of horizontal lines, following a simple numeration strategy which coded the locations from top-right to bottom-left (Figure 4b2). At the third stage of the game the latter strategy was combined with the use of two short vertical dashes indicating the new layer of the maze (Figure 4b3). These dashes were indistinguishable from those used to indicate the four locations of the first stage of the game. However, their meaning was different. While for first four locations they referred to vertices, for the newest location they meant "second" or perhaps "new". This shift could be interpreted in two ways: Either the dashes were bleached of their original meaning, becoming simply units for numeration, or they became polysemous, changing meaning depending on context. Considering that these interpretations do not affect the public meaning of the signals and that, as illustrated above, their private meanings might have varied between players, determining to what extent the unit "vertical dash" was meaningful is not a straightforward matter.

Figure 4. (a) Growth of the maze over the course of the game; (b) and (c) two different sign systems used to communicate the maze (the highlighted areas indicate developments in the sign systems, as described in Section 2.1.4)



(c)



The second sign system (Figure 4c) began at the first stage of the game with signals composed of horizontal lines which followed a simple numeration strategy to code the maze locations (Figure 4c1). At the second stage, three of the five new locations in the game were indicated by signals composed of horizontal dashes, again following a simple numeration strategy which coded the locations from top to bottom (Figure 4c2). At the third stage of the game the latter strategy was combined with the use of one horizontal line indicating the new layer of the maze (Figure 4c3). The line was indistinguishable from the ones used to indicate the four locations of the first stage of the game. However, the meaning was different. Whereas the lines used for the first four locations referred to specific maze locations, it is not clear what they meant for the newest location. Perhaps they had become almost meaningless units indicating mere otherness. Or perhaps they meant something like "on top of", indicating the new layer of the maze. Given that these interpretations do not affect the public meaning of the signals, determining in what way the unit "horizontal line" might be meaningful is again not straightforward.

3. Analysis of meaning in a semiotic referential game

The observations made in Section 2 suggest that ES can shed light on meaning in human communication systems. However, these observations are rather anecdotal in nature. If the same phenomena cannot be captured in a more principled way, then the light that ES can shed is a rather limited one. As noted in Section 1, semiotic referential games put greater constraints than coordination games on what participants have to communicate about; this provides an opportunity to examine meaning in a more controlled fashion. In the remainder of this paper we present a post hoc analysis of data from a referential communication study carried out by.

3.1. Description of the game

Figure 5. Screenshot from an early stage of game. The screen on the left was the sender's screen; the screen on the right was the receiver's

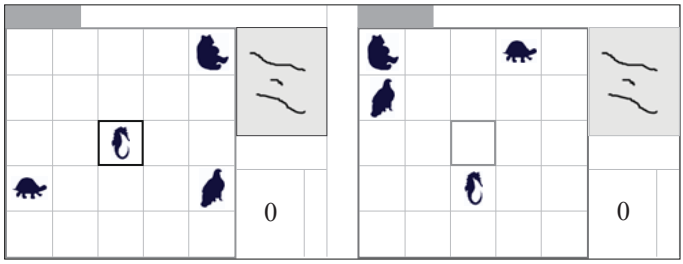
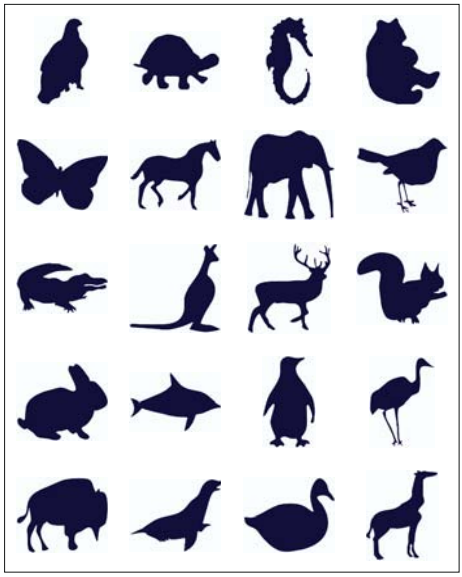


Figure 6. Referents used in the game. The top row shows the referents that were visible to players at the start of the game. After they had reached 75% success on these four, the next row of referents was added, and so on



Twelve dyads played a cooperative guessing game, sitting in separate locations with the same set of referents displayed in random locations in a 5x5 grid on a video monitor (Figure 5). The game consisted of a series of rounds. In each round, one player would play as sender and the other as receiver. The sender was informed of a target referent and had to convey information to the receiver that would help the receiver select the correct target referent on his or her own screen. If the receiver selected the correct target the round was counted as successful; if not, the round was counted as unsuccessful. Since the players played over the internet and were seated in separate locations, they could not speak to each other directly. Instead, the sender could convey information to the receiver, as in Galantucci's (2005) study, by making tracings on a digitizing pad with a magnetic stylus; these tracings were transformed in real time into on-screen signals such that the horizontal component of the tracings determined the horizontal component of the signal, but the vertical component of the tracing was replaced by a simple downward movement at a constant rate (Figure 1a). Players could not use this pad as an effective drawing or writing device (Figure 1b), even after prolonged practice, and to succeed at the task dyads had to cooperatively develop novel forms of communication (Galantucci, 2005). To help them in this, both players received feedback after each selection. Specifically, the receiver was shown what the target image had been and the sender was shown which image the receiver had selected. After the feedback phase, the next round began. Players swapped sender and receiver roles after each round.

Twenty silhouettes of animals (Figure 6) were used as referents, four of which appeared on the players' screens at the start of the game. The referents were presented as targets in a pseudo-random order: Dyads iterated through the four referents twice every eight rounds (in random order). A performance score was kept updated for each referent, based on the proportion of successful rounds in the cycle. If a dyad had at least 75% success on each of the four referents, the number of referents in the set was increased to eight, and the cycle length was increased accordingly to 16 rounds. The referent set and cycle length continued to be incremented in this way until either players had mastered all 20 referents or two hours of playing had elapsed.

Following the experiment, sign-sets were constructed for all players. A sign-set consisted of every referent for which the player in question had reached a 75% success rate paired with the last signal used to successfully communicate it. In Sections 3.2 and 3.3 we present analyses of these sign-sets aimed at investigating alignment between players and the meaningfulness of units smaller than a signal.

3.2. Private and public alignment

As observed in Sections 2.1.2 and 2.1.3, two or more individuals can use the same signal to communicate successfully (i.e., be publicly aligned), yet interpret

a signal in different ways (i.e., be privately misaligned). This phenomenon can be observed not only in laboratory languages and not only with respect to meaning. Wray and Grace (2007, 564) discussed data gathered by Fairman (2003) from letters written by semi-literate individuals in 19th Century England. The writers were native English speakers who can be assumed, in speech, to have been unexceptional in their pronunciation of such common phrases as "at home", "at all", and "take it" and to have been able to communicate these concepts unproblematically to other English speakers. In their letters, however, spellings such as *a tome*, *a torll*, and *taket* reveal that the way they parsed these phrases is strikingly at odds with the way a more literate Englishman would have done.[2]

As in this example, and as observed in Section 2.1.2, private misalignment may reveal itself publicly in certain circumstances. However, in many circumstances it will be entirely unobservable in individuals' public communicative behavior and will need to be elicited, as described in Section 2.1.3, in interviews with those individuals. This poses a problem for ES researchers that is considerably more acute than it is for researchers of real-world languages: Participants' memory of their own laboratory languages fades rapidly after the conclusion of the experiment, and any interviews need to be conducted immediately; interview questions that arise from analysis of the data (except to the extent that they can be generated automatically at the end of the study) cannot be asked. While Roberts and Galantucci's (2012) participants were able to use the signals in their sign-sets to communicate successfully, implying relatively good public alignment, the participants were not interviewed at all about their interpretation of the signals, making it impossible to investigate how well they were aligned privately. This does not mean that the data cannot be used to investigate such matters, however. While private and public alignment between the creators of a particular laboratory language is very much worth investigating, it is no less interesting to investigate alignment between new individuals exposed to the same language. There is, moreover, no limit to the number of new individuals who can be exposed to the same language (while the nature of the experiment restricts the number of creators to two), meaning that a small number of sign-sets can be used to gather a large number of data points. In Section 3.2.1 we describe such a study, in which we recruited sixteen judges (none of whom had been involved in the original study) to rate the iconicity of the signs produced in Roberts and Galantucci's (2012) study and measured the extent to which the judges were aligned with each other. We chose to focus on iconicity because it can be identified comparatively straightforwardly by exposing naïve participants to signals and referents. If an element of a signal can be identified as iconic, we can feel

[2] An interesting related finding comes from more recently gathered data from French speakers. In a study of gender acquisition by second-language learners of French, asked a control group of native speakers to assign gender to common nouns in French. She was surprised to find significant levels of disagreement (Harley, February 25, 2008).

relatively confident that it is to some extent meaningful. Iconicity is also an aspect of meaning that was measured for other purposes (and was thus known to vary) by Roberts and Galantucci (2012). We should not, however, restrict the search for meaningful structure to iconic meaning; in Section 3.3 we describe an approach to identifying meaningful structure that does not rely on the presence of iconicity.

3.2.1. Analysis

Our analysis was not the first time that the data in question had been shown to naïve judges. Roberts and Galantucci (2012) measured measured the *Transparency* of the sign-sets by asking judges to match referents with the signals that referred to them. While a Wilcoxon test indicated that judges did better than chance at matching up signs with referents ($W = 132$, $p < 0.001$), their success rate was not high. Overall, judges correctly matched 12% of signals with referents. Furthermore, there was no significant tendency for two judges rating the same sign-set to correctly match the *same* signal-referent pairs. This suggests low levels of alignment on the transparency of individual signs.

As a measure of the judges' private interpretation of the signs, however, this is very indirect. We investigated the question more directly by asking sixteen new judges explicitly to judge the iconicity of the signs in four of the sign-sets. The four sets chosen were the three sets with the highest Transparency scores ($Z = 1.35$, 1.75, and 2.5)[3] and one set chosen at random from those with the modal score (0.25). Each judge saw one player's signs from each of these four sign-sets (the order in which sets were presented was randomized, as was the order of signs within each set) and each player's sign-set was shown to eight judges.

The study proceeded as follows. First, the notion of iconicity was explained to the judges; in particular, it was stated that iconic signs resemble what they refer to, or resemble something closely connected with it. Examples were given of familiar iconic (e.g., a road sign and the word "cuckoo") and non-iconic signs (e.g., the word "cat"). The judges also gained an understanding of the game by playing a few rounds themselves (as both sender and receiver, with pictures of faces as referents). Then each judge was presented with a screen showing a player's signal (as a playable video, since signals in the game were displayed dynamically) and the referent it referred to. The judge's task was to answer yes or no to the question "Do you think the sign is iconic?" If the answer was yes, he or she was asked, "How is it iconic? What features of the sign correspond to what features of the animal?", to which an open answer was requested. If the judge answered no to the first question, a new sign appeared. Judges could take as long as they wished to answer the questions, but could not go back and change

[3] These scores are z-scores derived by taking the number of correct matches, subtracting from it the chance-level mean and dividing the result by the standard deviation. Since two judges matched signs for each player's sign-sets, the scores reported here are in fact mean values for the dyad.

their minds once they had submitted an answer. One judge's answers had to be discarded because he failed to understand the task, leaving fifteen sets of responses in total.

3.2.1.1. Closed-answer alignment

With respect to the first question ("Do you think the sign is iconic?") the judges' level of agreement could range from a 50–50 split (or a 4:4 ratio, where four judges thought the sign was iconic and four thought it was not) to full agreement (or a 0:8 ratio, where all judges thought either that the sign was iconic or that it was not). We therefore calculated a closed-answer alignment score by dividing the frequency of the most common response (yes or no) by the total number of responses. This produced a score between 0.5 (indicated a 50–50 split) and 1 (indicating unanimity). The results are shown in Table 1. The mean closed-answer alignment score for the four sets was 0.75 (SD = 0.01), meaning that on average judges were divided 2:6 on whether the signs were iconic. A Monte Carlo simulation, in which responses were generated at random 100,000 times, revealed that this level of alignment was greater than would be expected by chance ($p < 0.001$) and was not significantly greater for any one set than for any of the others. Overall, judges did not overwhelmingly find sign-sets to be iconic, however. For no set was the mean proportion of "yes" answers above 40%.

Table 1. Mean Transparency and Alignment scores for the four sign-sets

Dyad	Transparency	Closed-answer alignment	% answering "yes" to closed question	Open-answer alignment
1	1.75	0.75	31.20	0.39
2	2.50	0.74	32.90	0.46
3	0.25	0.75	30.00	0.36
4	1.35	0.77	39.60	0.38

3.2.1.2. Open-answer alignment

If participants answered yes to the question "Do you think the sign is iconic?" they were asked a second, open, question: "How is it iconic? What features of the sign correspond to what features of the animal?" This question produced

a variety of answers, such as "The two lines look like the two skinny legs of the animal" (Figure 7), or "Water or a beach was drawn, and that connects to where seals live" (Figure 8).

To calculate an open-answer alignment score the responses were assigned to the following 39 categories, which were based on the judges' responses: *antler(s), arm(s), back, beak, body, bottom, chest, claw(s), ear(s), feather(s), fin(s), foot/feet, footprint(s), habitat, fur/hair, hand(s)/paw(s), head, hug, indentation, leg(s), mouth, movement, neck, nose, pattern, point(s), pouch, scales, shape, shell, skin, span, stomach, tail, tentacle(s), tooth/teeth, trunk, waist, wing(s).* A few of these categories represent responses by one judge only, which were hard to fit into other categories. For example, one judge thought that signals representing the seahorse depicted the animal's "tentacles". Another judge said of a signal representing the squirrel that "it has a lot of points like this animal". The category *shape* was used for any answer that referred to the shape of the referent, without specifically mentioning any particular feature (such as the back or the legs). Legs and feet (as well as hands and arms) were distinguished because some judges distinguished between them in their responses. Some responses fell into two or more categories (e.g., "four legs and an antler were drawn", which was categorized under both *leg(s)* and *antler(s)*).

Figure 7. Signal representing a flamingo

Figure 8. Signal representing a seal

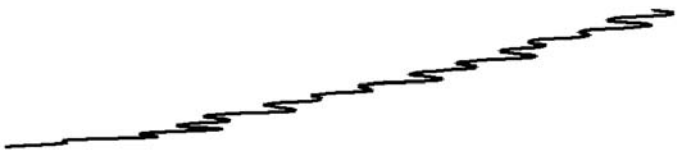
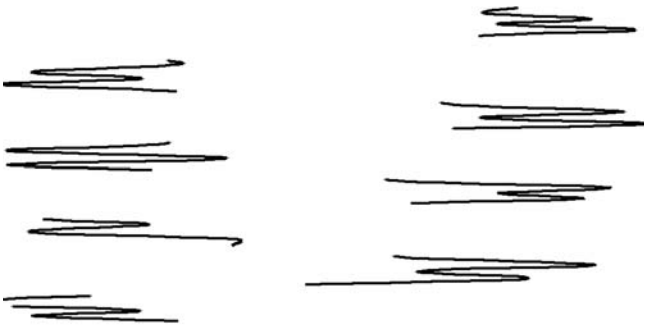


Figure 9. Signal representing an eagle

Figure 10. Signal representing a butterfly



Each sign was then given a score based on how well the judges' responses were aligned. If only one judge had rated a sign as iconic, it was discarded from the analysis; 33.8% of signs were discarded for this reason. If two or more judges considered a sign iconic, each judge's response was compared with the response of every other judge. Every time a judge's category appeared in another judge's response, it was counted as a hit. A hit meant that, if the two judges' responses were concatenated, the category in question would appear twice; each hit (H) was thus worth $\frac{2}{N_c}$, where N_c represents the total number of categories in the two responses. An open-answer alignment index (I_A) was then calculated for the sign, simply by dividing the sum of hit values by the total possible value:

$$I_A = \frac{\sum H}{\binom{N_j}{2}}$$

where N_j is the number of judges who found the sign iconic and therefore provided a response. This resulted in an index from 0 to 1, where 0 represented no alignment, and 1 represented complete alignment.

For example, four out of eight judges rated the signal for eagle in Figure 9 as iconic. Their responses fell into the following categories:

First judge:
beak,
Second judge:
head, body, leg(s),
Third judge:
body,
Fourth judge:
body, claw(s).

Only the category *body* appeared in more than one response, meaning that there were three hits in total. The open-answer alignment index for this signal was:

$$\frac{\frac{2}{4}+\frac{2}{5}+\frac{2}{3}}{\binom{4}{2}}=\frac{0.5+0.4+0.67}{6}=0.261$$

The butterfly signal in Figure 10, by contrast, received an alignment index of 1: All eight judges thought it represented the butterfly's wings, and nothing else.

The mean alignment index for each set can be seen in Table 1. A Monte Carlo simulation was run to test for significance. For each judge, each signal was reassigned the same number of categories as it originally received, but drawn at random (with replacement) from the full set of 39 categories. Alignment indices were then calculated. This was repeated 100,000 times. The number of times the mean alignment index was equal to or greater than that of the real data was then divided by 100,000 to calculate a p value. This revealed that the level of alignment for each of the four sets was greater than chance ($p < 0.001$). The open-question alignment scores should not be compared directly with the closed-question alignment scores. First, the two scores were by necessity calculated differently. Second, the open question applied only to those judges who had answered "yes" to the closed question. In other words, it is a measure of alignment between judges who all agreed that the sign was iconic. This is important because, in spite of this total agreement, they only agreed between 36% and 46% on *why* the sign was iconic. This is consistent with our observation that private misalignment need not lead to public misalignment. Although our judges had substantially different private interpretations of the iconic relationship between signal and referent (as shown by their responses to the open question), they agreed that the relationship *was* iconic, suggesting that—without adjusting their misaligned private interpretations—they could have used the signals to successfully communicate the referents to each other.

3.3. Meaningful and meaningless recombination

The vast majority of ES studies, regardless of category, are organized into a series of well defined short turns or rounds. As well as streamlining the dynamics of the experiment, this means that the communicative behavior of participants is conveniently broken into relatively cohesive units, or signs. Difficulties arise below the level of the sign, however. While it may be trivial to identify a particular signal as referring to an eagle, it is typically non-trivial to identify a part of that signal as referring to a particular feature of the eagle. Roberts and Galantucci (2012) measured the *combinatoriality* of their participants' sign-sets: that is, the extent to which the signs were composed of recurrent meaningless units, as in natural-language phonology or non-ideographic writing systems. Combinatoriality

should be distinguished from *compositionality*, in which meaningful units are recombined, as in morphology or syntax. (The degree of compositionality in the system would also have been an interesting question, but was not relevant to the question the authors were addressing and was thus not measured.) Roberts and Galantucci (2012) directly measured the degree to which parts of signals (*forms*) recurred between signals. They did not, however, attempt in any principled way to distinguish between meaningful and meaningless forms, noting that:

> Communication systems are more wasteful with their meaningful units than with their meaningless ones: There are considerably more of them, and they are recombined less. [Our measure] should therefore strongly correlate with true [i.e., meaningless] combinatoriality. The algorithm cannot be used, however, to identify the meaningfulness of an individual stroke.

In Section 3.3.1 we will examine methods that might allow such distinctions to be made.[4]

3.3.1. Analysis

To identify meaningful recurrence of units, or compositionality, three steps are necessary (cf. Goldin-Meadow, Mylander, & Butcher, 1995):

1. identifying subunits of signals (henceforth *forms*);
2. categorizing referents based on shared features;
3. scoring correspondence of forms to meaning categories.

In English, for example, the form /bəri/ (a subunit of the word spelled *raspberry*) corresponds relatively well to the meaning category fruit. The form /riz/, by contrast, occurs in several of English words (such as *breeze*, *trees*, and *raspberries*), yet does not correspond very reliably to any particular meaning.

We applied these steps to our own data as follows.

3.3.1.1. Identifying forms

We used the same method as Roberts and Galantucci (2012) to identify units within the participants' signals. First we used an algorithm to break up the signals into forms—strokes of the stylus separated by space. We then used a second algorithm to compare forms between signals and identify whether or not they were equivalent (see Roberts & Galantucci, 2012, 316–318, for a full description

[4] It should be borne in mind that drawing a sharp distinction between combinatoriality and compositionality, or between phonology and morphology, is not always possible. Phonaesthemes in spoken languages straddle the boundary between meaningless and meaningful. The phoneme inventories of sign languages, moreover, often include units that were iconic, and thus meaningful, in origin; residual iconicity may well remain for very long periods. As suggested in this paper, the answer to the question of whether a given unit is meaningful may depend on who is asked (as well as when or how).

of these algorithms). This resulted in a set of *unique forms* for each sign-set. Because we were interested specifically in forms that could be recombined, we discarded those forms that were used to refer to only one referent.

3.3.1.2. Categorizing referents

We categorized the referents according to the features chosen by the sixteen judges in Section 3.2.1. Features that could apply to all referents (e.g., *body* or *shape*), applied to only one referent (e.g., *antlers*, or *hug*), or were coextensive with another feature (e.g., *feet*, which was coextensive with *leg*) were ignored. This left ten features with which we constructed ten meaning categories. That is, we categorized all the animals according to whether or not they exhibited the following features: *legs, wings, ears, tail, beak, nose, scales, water*,[5] *fur*, and *claws*. The referents *bear, horse, kangaroo, squirrel, deer, rabbit, buffalo*, and *giraffe* were included in the category *fur*, for instance.

3.3.1.3. Scoring correspondence

For every sign-set we paired every meaning category with every unique form and scored the pairing by dividing the number of shared referents (those that both were included in the category and were referred to using the form) by the number of shared referents plus the number of unshared referents (those to which the form referred, but were not included in the category, or vice versa). This produced a correspondence index from 0 to 1, in which 0 meant "no correspondence" and 1 meant "complete correspondence". For example, consider the category *wings*, which contained the referents *eagle, bird, butterfly, penguin, flamingo*, and *duck*. With respect to this category, a form used to represent the referents *eagle, bird, penguin, flamingo, duck* and *bear*, would receive a correspondence score of $5/(5+2) = 5/7 = 0.71$, since there would be five shared referents (*eagle, bird, penguin, flamingo*, and *duck*, all of which have wings and whose signals contained the form in question) and two unshared (*bear*, which does not have wings, and *butterfly*, whose signal does not contain the form).

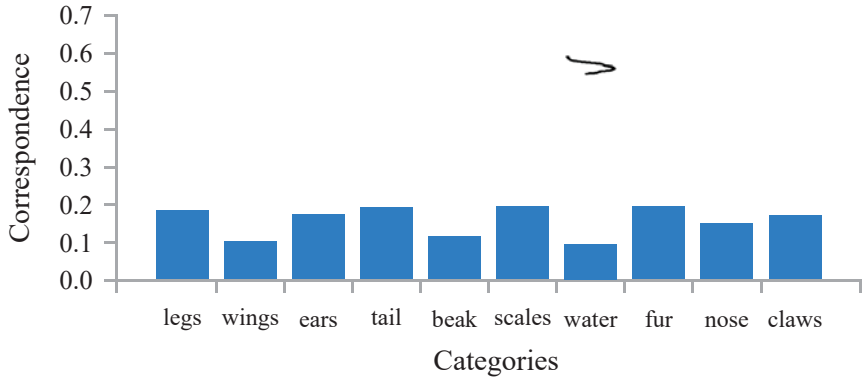
The overall mean correspondence score for all forms and categories was 0.12, although scores ranged from 0 to 0.75. As a measure of meaningfulness, correspondence scores should be treated with caution, however. The mean score is particularly misleading. The most straightforwardly meaningful form conceivable (i.e., one that corresponded perfectly with one meaning category, and not at all with any other) would have a correspondence score of 1 for a particular category, but a *mean* correspondence score of only 0.1 (i.e., 1 divided by the number of categories). On the other hand, while an high correspondence score between a single form and a single category might suggest a good fit between the two, it must not be considered in isolation. A form might co-occur very reliably with

[5] The feature water was used instead of habitat, since only two habitats were mentioned by the judges: water and racetracks, the latter of which applies only to horses.

every meaning category, resulting in high correspondence scores; such a form, however, would either be meaningless or would mean something like "animal", which for the referent set in question would be as good as meaningless. To use the correspondence scores to get at meaningfulness, their distributions must be taken into account.

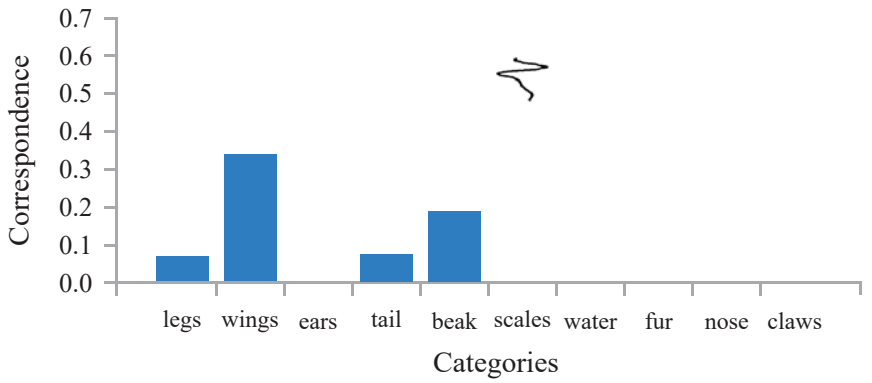
A meaningful form is one that corresponds relatively well to *a small proportion* of the available meaning categories—where correspondence scores are relatively unequally distributed, in other words. It follows that a measure of whether or not a form is meaningful should be obtainable using an index of inequality.[6] We therefore calculated the Gini coefficient for each form. The Gini coefficient was devised by Corrado Gini, who defined it as the "the mean difference from all observed quantities" (see Ceriani & Verme, 2012, for a more detailed account). It is particularly well known for its use by economists to calculate income inequality and can be calculated in a number of different ways (Abounoori & McCloughan, 2003; Milanovic, 1994, 1997). However, it is usually defined based on the Lorenz curve, a plot of the proportion of the total income of a population (y) that is cumulatively earned by the bottom x% of the population. On such a plot, a line at 45° represents perfect equality and the Gini coefficient can be calculated by dividing the area between the 45° and the Lorenz curve by the total area under the 45° line. For our purposes, a form with a high Gini coefficient (i.e., where form-category correspondence is relatively unevenly distributed) is relatively likely to be meaningful. The form in Figure 11, for example, has a Gini coefficient of 0.12 and is far less likely to be meaningful than the form in Figure 12, with a Gini coefficient of 0.75.

Figure 11. Form with Gini coefficient of 0.12



[6] Tamariz and Smith (2008) devised a measure called *RegMap* to do a similar task, namely to measure the regularity in mappings between signals and meanings. However, this measure was designed for cases where there are clear meaning dimensions with multiple values (such as color). This is not the case for our dataset.

Figure 12: Form with Gini coefficient of 0.75



It is important to note that the Gini coefficient is a measure of *relative* distribution, and is not sensitive to the particular values of the form-category correspondences. Indeed, the form in Figure 12, with its relatively high Gini coefficient of 0.75, does not have a correspondence score of more than 0.33 with any meaning category. In other words, the form is rated as meaningful, while not corresponding especially well to any meaning! This is not the serious problem it may appear. A meaningful form should not be expected to correspond perfectly to any meaning category; the English prefix *un-*, for example, is very clearly meaningful, yet does not occur in all negative adjectives. Such cases can very easily arise in several different circumstances, such as when different units are used to convey the same meaning (e.g., un-, in-, non- etc.), when meaningful forms are sometimes used meaninglessly (as with the syllable /mu/ in English, which refers iconically to the sound a cow makes, but also occurs in many unrelated words), or when meaningful forms are used for different meanings (as described in Section 2.1.4). Conversely, meaningless forms might also be reinterpreted as meaningful (cf. Kirby et al., 2008). The Gini coefficient may therefore be particularly well suited to datasets in which meaningful structure is only just emerging, particularly as a measure of *relative* meaningfulness. However, it should be borne in mind that it may overestimate *absolute* meaningfulness in our dataset. The overall mean Gini coefficient for all of Roberts and Galantucci's (2012) sign-sets was 0.46 (SD = 0.086), which cannot be taken to imply that the combinatoriality measured by Roberts and Galantucci (2012) was overwhelmingly meaningful. We performed two further analyses. First, we correlated the mean Gini values for each sign-set with the the Transparency scores, but found no significant relationship; second, to see whether meaningfulness increased or decreased over time in our datasets, we took forms from different stages in each game (after the dyad had reached 75% on

the first four referents, on the first eight referents, and so on) and performed the same analysis on each set. However, there was no clear trend in either direction. This does not mean our approach is useless; it is quite possible that the measure is picking up on meaningful structure that our judges did not consider iconic. A better approach to validation would be to compare our approach directly with self reports by the participants who constructed the system. Such reports were not gathered for the dataset examined here, and doing so is not trivial—there is a serious danger of participants forgetting real patterns and creating false ones post hoc—but, if done carefully, this might be a fruitful focus for a future study, as would an approach applying the measure to natural language. What is to be gained is a broadly applicable measure for distinguishing meaningful from meaningless structure.

4. Conclusion

In this paper we discussed ways in which different kinds of ES study can shed light on meaning in emergent communication systems. We focused in particular on two observations: that success in communicating (public alignment) need not imply that the individuals interpret the communicative signals in the same way (private alignment) and that distinguishing between meaningful and meaningless combinatorial units can be far from straightforward. We supported these observations, drawn from semiotic coordination games carried out by Galantucci (2005) and Galantucci and Roberts (2012), with analyses of a dataset produced by Roberts and Galantucci (2012). As well as providing support for the earlier anecdotal observations, our analyses illustrated how ES researchers can measure the meaningfulness of combinatorial forms and the degree of public and private alignment between different users of the same laboratory language (even if the original creators of the language are no longer available).

We feel we have also illustrated something else: that analyzing meaning in Experimental Semiotics is a difficult process. It is tempting to attribute this difficulty to the nature of the communication systems involved. It is hard, for instance, to detect systematic properties in a system in which such properties may only just be emerging, and where the boundary between inclusion and non-inclusion in the system for a given element is fuzzy. Even if that boundary is clear, laboratory languages tend to be small, and it is hard to detect systematic properties when there are few instances to which the system applies (cf. Roberts and Galantucci, 2012, 312). Furthermore, it is hard to gain insight by questioning users of the communication system if those users' memory of the system fades within hours of using it. Yet these problems should not all be laid at the door of Experimental Semiotics. Even for real-world languages it can be hard to get reliable measures by questioning speakers; as Wray and Grace (2007, 544) noted:

> Lecturers have done well if they get through a syntax class without someone questioning their allocation of asterisks, even when the grammaticality judgements are supposed to be universal. Linguists also know that it is not a good idea to ask members of the general public to judge complex sentences for grammaticality, because they find it difficult to come up with the responses predicted by the theory.

Nor is it always straightforward to find the boundary between meaningful and meaningless elements in real-world languages. Phonological analyses of sign languages, for example, are constrained by the fact that iconic meaning is so pervasive; Sandler et al. (2011) argued that in ABSL phonological structure is only just beginning to emerge. Perniss, Thompson, and Vigliocco (2010) demonstrated moreover that, even in spoken languages, iconicity is far more pervasive than previously thought (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016), while Ladd (2012) argued that the very notion of duality of patterning is considerably more complex and problematic than it appears on first sight. And while identifying forms and meaning-categories may be a challenge for experimental semioticians, it turns out that defining what constitutes a word, in a cross-linguistically useful sense, is surprisingly challenging even outside the lab (Dixon and Aikhenwald, 2002).

If these matters have appeared straightforward, it is likely because our perspective has been skewed by a tendency to focus too much attention on too small a sample of languages (Wray and Grace, 2007, 546–548). By analyzing the communicative behavior of individuals prevented from even using a pre-existing language, and by forcing researchers to rethink long-established notions, we hope that Experimental Semiotics can help reduce that skew and shed useful light on meaning (among other features of language and communication). We hope further that we have shed light in this paper of how precisely one might go about doing so.

Acknowledgements

We are grateful to Christian Kroos, Carrie Theisen and Theo Rhodes for their contributions to producing the dataset analyzed in this paper. We also gratefully acknowledge the support of the National Science Foundation (BCS-1026943).

References

Abounoori, E. & McCloughan, P. (2003). A Simple Way to Calculate the Gini Coefficient for Grouped as Well as Ungrouped Data. *Applied Economics Letters*, *10* (8), 505–509.

Ayoun, D. (2007). The acquisition of grammatical gender in L2 French. In D. Ayoun (Ed.), *French Applied Linguistics* (pp. 130–170). Amsterdam: John Benjamins.

Blasi, D.E., Wichmann, S., Hammarström, H., Stadler, P.F., & Christiansen, M.H. (2016). Sound-meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113* (39), 10818–10823.

Ceriani, L. & Verme, P. (2012). The origins of the Gini index: Extracts from Variabilità e Mutibilità. *Journal of Economic Inequality*, *10* (3), 421–443.

Cornish, H., Christiansen, M.H., & Kirby, S. (2010). The emergence of structure from sequence memory constraints in cultural transmission. In A.D.M. Smith, M. Schouwstra, B. De Boer, & K. Smith (Eds.), *The Evolution of Language: Proceedings of the 8th International Conference (EVOLANG8)* (pp. 387–388). Singapore: World Scientific.

Dingemanse, M., Blasi, D.E., Lupyan, G., Christiansen, M.H., & Monaghan, P. (2015). Arbitrariness, iconicity and systematicity in language. *Trends in Cognitive Sciences*, *19* (10), 603–615.

Dixon, R.M.W. & Aikhenwald, A.Y. (2002). *Word: A Cross-Linguistic Typology*. Cambridge: Cambridge University Press.

Fairman, T. (2003). Letters of the English labouring classes and the English language. In M. Dossena & C. Jones (Eds.), *Insights into Late Modern English* (pp. 265–282). Bern: Peter Lang.

Galantucci, B. (2005). An experimental study of the emergence of human communication Systems. *Cognitive Science*, *29* (5), 737–767.

Galantucci, B. (2009). Experimental Semiotics: A new approach for studying communication as a form of joint action. *Topics in Cognitive Science*, *1* (2), 393–410.

Galantucci, B., Fowler, C.A., & Richardson, M.J. (2003). Experimental investigations of the emergence of communication procedures. In R. Sheena & J. Effken (Eds.), *Studies in Perception and Action VII – Proceedings of the 12th International Conference on Perception & Action (ICPA)* (pp. 120–124). Mahwah, NJ: Lawrence Erlbaum Associates.

Galantucci, B. & Garrod, S. (2010). Experimental Semiotics: A new approach for studying the emergence and the evolution of human communication. *Interaction Studies*, *11* (1), 1–13.

Galantucci, B., Garrod, S., & Roberts, G. (2012). Experimental Semiotics. *Language and Linguistics Compass*, *6* (8), 477–493.

Galantucci, B. & Roberts, G. (2012). Experimental Semiotics: An engine of discovery for understanding human communication. *Advances in Complex Systems*, *15* (3–4), 1150026: 1–13.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, *31* (6), 961–987.

Goldin-Meadow, S., Mylander, C., & Butcher, C. (1995) .The resilience of combinatorial structure at the word level: Morphology in self-styled gesture systems. *Cognition*, *56* (3), 195–262.

Harley, H. (February 25, 2008). "You say feminine, I say masculine, let's call the whole thing off" (Blog Post) Retrieved on Jan 3, 2016, from http://itre.cis.upenn.edu/~myl/languagelog/archives/005411.html

Hurford, J.R. (1989). Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua*, *77* (2), 187–222.

Kirby, S., Cornish H., & 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*, *105* (31), 10681–10686.

Ladd, D.R. (2012). What Is duality of patterning, anyway? *Language and Cognition*, *4* (4), 261–273.

Milanovic, B. (1994). The Gini-type functions: An alternative derivation. *Bulletin of Economic Research*, *46* (1), 81–90.

Milanovic, B. (1997). A simple way to calculate the Gini coefficient, and some implications. *Economics Letters*, *56* (1), 45–49.

Perniss, P., Thompson, R.L., & Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, *1*, 1–15.

Roberts, G. (2010). An experimental study of the role of social selection and frequency of interaction in linguistic diversity. *Interaction Studies. 11* (10), 138–159.

Roberts, G. & Galantucci, B. (2012). The emergence of duality of patterning: Insights from the laboratory. *Language and Cognition*, *4* (4), 297–318.

Sandler, W., Aronoff, M., Meir, I., & Padden, C. (2011). The gradual emergence of phonological form in a new language. *Natural Language and Linguistic Theory*, *29* (2), 503–543.

Scott-Phillips, T.C., Kirby, S., & Ritchie, G.R.S. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113* (2), 226–233.

Tamariz, M. & Smith, A.D.M. (2008) Regularity in mappings between signals and meanings. In A.D.M. Smith, K. Smith, & R.F. Cancho (Eds.), *The Evolution of Language: Proceedings of the 7th International Conference (EVOLANG7)* (pp. 315–322). Singapore: World Scientific.

Theisen, C.A., Oberlander, J., & Kirby, S. (2010). Systematicity and arbitrariness in novel communication systems. *Interaction Studies*, *11* (1), 14–32.

Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and Cognition*, *4* (4), 357–380.

Wray, A. & Grace, G.W. (2007). The Consequences of Talking to Strangers: Evolutionary Corollaries of Socio-Cultural Influences on Linguistic Form. *Lingua*, *117* (3), 543–578.