# Evolving artificial sign languages in the lab: from improvised gesture to systematic sign

Yasamin Motamedi[a,b,*], Marieke Schouwstra[b], Kenny Smith[b], Jennifer Culbertson[b], Simon Kirby[b]

[a]*Department of Experimental Psychology, University College London, 26 Bedford Way, London, UK, WC1H 0AP*
[b]*Centre for Language Evolution, University of Edinburgh, 3 Charles Street, Edinburgh, UK, EH3 9AD*

**Abstract**

Recent work on emerging sign languages provides evidence for how key properties of linguistic systems are created. Here we use laboratory experiments to investigate the contribution of two specific mechanisms—interaction and transmission—to the emergence of a manual communication system in silent gesturers. We show that the combined effects of these mechanisms, rather than either alone, maintain communicative efficiency, and lead to a gradual increase of regularity and systematic structure. The gestures initially produced by participants are unsystematic and resemble pantomime, but come to develop key language-like properties similar to those documented in newly emerging sign systems.

*Keywords:* silent gesture, iterated learning, interaction, transmission, sign language, language evolution

## 1. Introduction

Languages exhibit systematicity; single utterances are not isolated, independent units but form part of a structured system of interdependent elements. We see systematicity across levels of language, in the lexicon, morphology and syntax. Parts of signals are re-used and recombined

*Corresponding author
*Email address:* y.motamedi@ucl.ac.uk (Yasamin Motamedi)

across utterances and correspond systematically to different aspects of the meanings being conveyed. For example, the noun phrases *blue shoes* and *red shoes* both include *shoes*, indicating which part of their meaning is shared, and differ on their descriptors, *blue* and *red*, this difference in form signalling a difference in meaning. The prevalence of systematic structure of this kind across languages and modalities points to its status as a fundamental property of language. One of the central challenges of language evolution research is to determine the mechanisms through which systematicity arises. Cultural evolutionary accounts propose that it develops as language adapts to pressures arising from language use and the transmission of language to new learners (Raviv & Arnon, 2018; Raviv et al., 2019; Giudice, 2012; Beckner et al., 2017; Kirby et al., 2008, 2015; Cornish et al., 2013; Silvey et al., 2014). Investigation of such cultural processes thus requires the observation of communication systems at different stages of linguistic emergence. Currently, there are two main sources of evidence available for such observations: emerging sign systems, providing data from natural languages which are in the early stages of developing linguistic structure, and experimental research modelling language early in its evolutionary development. Here, we combine these approaches, observing the emergence of manual communication systems in the laboratory. We focus on how the cultural mechanisms of interaction and transmission drive the evolution of these systems. In essence, we create a controlled environment in which we can observe the evolution of miniature artificial sign languages.

### *1.1. Field research: Homesign and emerging languages*

Observations from homesign (Goldin-Meadow, 2003; Haviland, 2013) and emerging sign languages (Aronoff et al., 2005; Senghas et al., 2004; De Vos, 2014) provide the only naturally occurring evidence of language at its earliest stages, and give us crucial insights into the different cultural contexts and mechanisms that affect the structure found in communicative systems.

Homesign systems are created by deaf children who are not exposed to an accessible conventional language early in their development (usually being born to hearing parents), and must improvise ways to communicate. The structural properties of these systems do not appear to directly reflect parental input, either from spoken language or infant-directed gestures (Goldin-Meadow & Mylander, 1983, 1998). Homesigns do exhibit some of the structural features of established languages, such as regularities in syntax (Goldin-Meadow & Feldman, 1977), morphology (Goldin-Meadow et al., 1995) and lexical categorisation (Goldin-Meadow et al., 1994; Haviland, 2013). However, they also differ from established languages in a number of ways. They show less systematic structure and less regularity than sign languages (Goldin-Meadow et al., 2014), and lack consistency across users of a single system (Richie et al., 2014). The use of homesigns in communication is often limited and asymmetrical: homesign systems are developed by deaf individuals and encapsulated within their family. While hearing family members may use the system to some extent, only the deaf family members use homesigns as a primary communication system. Further, homesign systems typically persist for only a single generation; lack of a community halts further transmission, and thus evolution, of the system.

Observation of the development of early sign languages has also illuminated the potentially critical roles of interaction between users in a community and transmission to new community members. Emerging sign languages arise when communities are formed by deaf individuals, who lack a conventional language model, or who are otherwise cut off from pre-existing languages. In many cases of sign language emergence, linguistic systems begin as improvised homesigns within family units and develop as they are learnt by, and transmitted to, a wider community. These communities may emerge due to high rates of hereditary deafness, as is the case with many village sign languages such as Al-Sayyid Bedouin Sign Language in Israel (Aronoff et al., 2005), Kata Kolok in Indonesia (De Vos, 2014) and Adamarobe Sign Language in Ghana (Nyst, 2010), or they may emerge due to changes in educational policy, such as the provision of schools for the deaf. The latter led to the development of Nicaraguan Sign Language (NSL), which emerged in the late 1970s after a deaf school was established in Managua, and deaf individuals who had developed their own homesign systems were then able to interact with each other and develop a conventionalised language within the school (Kegl et al., 1999; Senghas & Coppola, 2001; Senghas et al., 2004). Studies of emerging sign languages have shown how linguistic features such as conventional word order (Sandler et al., 2005) and role shift (where signers take on the role of another 'character' in the discourse; Kocab et al., 2014) emerge and change over time. Senghas et al. (2004) demonstrate how motion events that were signed holistically (conflating manner and path) in the first cohort of NSL signers became sequential (separating manner and path signals) in later cohorts. Their results are particularly surprising, as simultaneous structure is a common modality-specific property of sign language morphology, which allows iconic event representation (Aronoff et al., 2005; Senghas et al., 2004); as such, it is striking that new learners of NSL do not exploit this iconicity.

Goldin-Meadow et al. (2014) analyse the consistency of handshape for nominals and predicates in four groups of signers: Nicaraguan homesigners, signers of NSL cohort 1, signers of NSL cohort 2, and signers of American Sign Language (ASL). Their results suggest that ASL, the oldest and most stable language in the sample, exhibits the most consistent handshapes across signers. By contrast, homesigners exhibited low consistency in handshapes used across the group, as would be expected from individual innovators. The NSL cohorts illustrate the link between these initial inconsistent systems and later systematicity, becoming increasingly consistent in their handshapes from cohort 1 to 2. These results suggest the importance of a community in the development of a language. Signs conventionalise and become more regular through use within the community; this process is further entrenched through transmission to new community members.

### *1.2. Experimental research: communication, iteration and gesture*

The early sign systems described above provide a valuable perspective on language emergence, allowing us to generate hypotheses about the mechanisms that drive the emergence of systematic structure in language. These systems are thus crucial to our understanding of language evolution. However, natural language observation requires large-scale longitudinal study of phenomena generally outside the researcher's control. Experimental research, on the other hand, presents the opportunity to test our hypotheses by manipulating both the linguistic structures and the social environments we wish to investigate.

Previous work on experimental semiotics has investigated how interaction shapes communication. Participants in these studies take part in communication games in which they draw concepts for other participants who attempt to interpret them (similar to *Pictionary*; Healey et al., 2007; Garrod et al., 2007; Fay et al., 2010). Findings suggest that repeated interaction leads to an increase in the production of symbolic signals: drawings become smaller, less iconic and less complex, leading to more efficient (i.e., faster) and more successful communication.

Alongside this, studies using the iterated learning paradigm have tested how transmission of a system to new learners affects linguistic structure (Kirby et al., 2008). In an iterated learning experiment, an initial set of participants is trained on an artificial language and then asked to reproduce it. Their reproductions are then passed on as the input for the next participant, who in turn attempts to reproduce what they have learnt. This process is repeated, with the output from one participant used as input for the next, modelling a process analogous to the generational transmission of language. Results show that systematic structure emerges through this process, leading to the claim that structure develops in response to a pressure for systems to be learnable (Beckner et al., 2017; Kirby et al., 2008, 2015, 2014; Carr et al., 2016).

More recently, several studies have probed how interaction and transmission may work together to shape linguistic structure. Kirby et al. (2015) and Carr et al. (2016) compare systems created via interaction alone with those which emerge through a combination of interaction and transmission. For the former, participants communicated in closed pairs, using a new language repeatedly with only each other. For the latter, chains of pairs were created, with one pair learning from the system produced by a previous pair (as in the iterated learning paradigm described above). Signals produced by chains of pairs evolved to be both learnable and useful for communication, while signals produced by closed pairs became useful for communication, but lacked the kind of structure which would make them easily learnable (e.g., compositionality). Similar results have been found in the graphical medium (Theisen-White et al., 2011; Caldwell & Smith, 2012; Fay & Ellison, 2013). For example, Theisen-White et al. (2011) had pairs of participants convey concepts which shared either a thematic feature (such as 'doctor' and 'hospital') or an entity type (such as 'hospital' and 'school'). Pairs of participants who communicated in a closed group introduced some systematic structure in their signals. For example, participants might re-use a (thematic) stethoscope across their drawings for 'doctor' and 'hospital'. However, only transmission of the signals to new pairs led to a cumulative increase in systematic structure over generations. Kirby et al. (2015) propose that structural differences in the systems produced by transmission and interaction reflect specific pressures each context brings: in their study, compressibility in the case of transmission and expressivity in the case of interaction. This does not mean that we would expect systems borne out through interaction alone to be wholly unlearnable —they must still be used by human minds and bodies, after all— but that these systems are a product of the trade off between these competing pressures. Indeed, experimental studies focussing on communication alone have found that compositionality can arise without transmission, when a pressure for compressibility is otherwise introduced (Raviv et al., 2019; Winters et al., 2018; Nölle et al., 2018). For example, Raviv et al. (2019) introduced a pressure for compressibility by using an expanding meaning space, such that the increase in the number of meanings participants had to communicate meant that participants were forced to generalise their existing system to convey new meanings. A system lacking in systematic structure would not allow this kind of generalisation, as well as becoming increasingly unwieldy as the size of the system increased. In this case, the authors found a cumulative increase in compositionality over rounds of communication.

As mentioned above, here we are interested in further understanding the trade-off between interactive and transmissive pressures, by investigating how they work in combination, and in isolation. In particular, we model these processes in the emergence of manual communication systems. To do this, we capitalise on the silent gesture paradigm. Silent gesture studies ask hearing participants with no knowledge of any sign language to communicate using gesture but no speech. Previous work suggests that this paradigm may reduce the influence of prior linguistic knowledge on participants' behaviour, revealing the cognitive biases that might shape linguistic systems. For example, Goldin-Meadow et al. (2008), amongst others (Hall et al., 2013; Schouwstra & de Swart, 2014) showed that participants from different linguistic backgrounds overwhelmingly produce Subject-Object-Verb (SOV) word order when gesturing, favouring structures which are most common cross-linguistically. Smith et al. (2017a) combine silent gesture with iterated learning to investigate how the expression of motion events develops via transmission of emerging artificial sign systems. This study offers an experimental comparison with the work on motion events in NSL described in section 1.1. Recall that signs for motion events using segmented structure (as compared to earlier simultaneous signs) emerged in the second NSL cohort (Senghas et al., 2004). In Smith et al. (2017a), participants produced gestures for videos showing a ball moving in varying manners (e.g. bouncing or rolling) and along different paths (e.g. a slope or a circular path). The gestures individual participants produced were then used to train further participants, who then produced gestures for the same set of scenes. As the gestural systems were transmitted, they became increasingly systematic. Moreover, although simultaneous structures were favoured, some sequential structures did emerge. These could be amplified by the transmission process, becoming frequent in a particular chain, or be weeded out as simultaneous structures took over. This work points to silent gesture as a method which can be combined with iterated learning to shed light on how individual behaviours may persist and be amplified depending on the pressures at play.

### *1.3. From pantomime to sign language*

We build on the body of literature discussed above to investigate the evolution of artificial manual communication from initial unstructured pantomime to sign language-like systems. We use silent gesture combined with iterated learning to understand the effects of interaction and transmission on the emergence of systematic structure over time. Use of the manual modality both minimises interference from participants' existing linguistic system, and allows us to compare our findings with the main source of emerging systems in natural environments: manual systems such as homesign and new sign languages (see section 1.1).

Hearing participants without knowledge of any sign language were asked to communicate about a set of concepts using only gesture. Participants across the set of experiments were asked to learn from gestures produced by previous participants, and then asked to either communicate with a partner using only gesture, or produce gestures for the set of concepts without a partner. Following Theisen et al. (2010) and Theisen-White et al. (2011), the concepts participants were asked to communicate about comprised a structured set differing on two dimensions (thematic and functional). Across a series of three experiments, we explicitly test the two mechanisms of interest, interaction and transmission, in isolation and in combination with each other.

We predict that the earliest, pantomime-like stages in the evolution of manual communication will feature low systematicity and high iconicity. Signals will be used unsystematically and independently of each other; the form of signals referencing similar meanings will not necessarily use similar properties. Relatedly, because gestures do not yet form part of a conventionalised code, productions are expected to differ both within an individual and across individuals communicating with each other (Klima & Bellugi, 1979; McNeill, 2000). By contrast, the later stages of evolution should more closely resemble sign languages: coded, conventionalised systems used by a community. Linguistic signs are predicted to become more efficient in form, for example, showing fewer redundancies, or requiring less physical effort to produce (Fay & Ellison, 2013; Caldwell & Smith, 2012; Gibson et al., Forthcoming), and signals should be used systematically, being re-used and re-combined consistently across the system to predictably differentiate between referents (Kirby et al., 2008). Importantly, we predict that the extent to which these more language-like features emerge will be dependent on both interaction and transmission being present.

In experiment 1, we show that gestures produced without a model evolve to become more systematic as they are transmitted to new learners. In experiment 2, we confirm the crucial role of both mechanisms, showing that transmission alone leads to learnable gestures, while interaction alone leads to communicatively efficient gestures. Finally, in experiment 3, we verify that an explicit pressure for efficient communication (present in experiments 1 and 2) is not necessary for more efficient gestures to develop; rather, it is embedded in the act of interaction itself.

## 2. Experiment 1: interaction and transmission

### *2.1. Methods and Materials*

In experiment 1, a first set of participants served as seed participants (described in section 2.1.3), who recorded a single gesture for a single concept in the meaning space. Following collection of these seed gestures, participants took part in the main experiment in pairs, organised into transmission chains of five generations. The gestures recorded in the seed collection stage were used as initial training gestures for the transmission chains in the main experiment.

#### *2.1.1. Participants*

48 seed participants (aged 18 to 33, median age 22) were recruited to record an initial set of videos, each participant recording a single video. Seed participants were unpaid volunteers recruited from the undergraduate and postgraduate population at the University of Edinburgh. The

seed task took no more than 5 minutes to complete. A further 50 participants (aged 18 to 32, median age 20) were recruited for the main experiment, from the University of Edinburgh careers website. All participants were self-reported right-handed native English speakers, with no knowledge of sign language. The main experiment took approximately 1 hour to complete, and participants were paid £7 for participation.

### *2.1.2. Materials*

Participants were presented with items from a meaning space containing a total of 24 meanings (see figure 1). Following Theisen-White et al. (2011), items in the meaning space shared an association either on the thematic dimension (for example, *chef* and *restaurant* share the thematic dimension of cooking), or on the functional dimension (for example, *chef* and *photographer* share the functional dimension of person). Meanings were presented as orthographic words in order to avoid ambiguity.

| | | **Functional dimension** | | | |
|---|---|---|---|---|---|
| | | **person** | **location** | **object** | **action** |
| **Thematic dimension** | **food** | chef | restaurant | frying pan | to cook |
| | **religion** | vicar | church | bible | to preach |
| | **photography** | photographer | darkroom | camera | to take a photo |
| | **music** | singer | concert hall | microphone | to sing |
| | **hair styling** | hairdresser | hair salon | scissors | to give a haircut |
| | **law enforcement** | police officer | prison | handcuffs | to make an arrest |

Figure 1: The meaning space. Concepts shown in the rows of the table share the thematic association of particular professions, while items in columns share the functional associations of *person*, *location*, *object* and *action*.

Participants (including seed participants) were placed in individual experiment booths, in front of an Apple Thunderbolt monitor with an affixed Logitech webcam. Both monitor and webcam were connected to an Apple Macbook Air laptop running Psychopy (Peirce, 2007) and

VideoBox (Kirby, 2016), custom software developed to record and stream video between networked computers.

### 2.1.3. Procedure: seed collection

Our initial seed participants were asked to produce a gesture for a single concept from the meaning space. The collection of seed videos thus provided a set of unique gestures used as the first training set for the transmission chains. Participants were told that they would be presented with a concept, and that they should communicate that concept using only gesture. Participants were instructed that they should not speak whilst doing the task, that they should not attempt to manually spell concepts, and that they should remain seated throughout the task. No audio was recorded at any stage of the experiment. All participants were given one practice trial in which they were presented with the concept *angry*, and asked to communicate this concept using only gesture. The target concept was shown onscreen and participants could begin to record their gesture by pressing the space bar. After pressing the space bar, participants were given a 3 second countdown on screen before recording started. During recording, participants were shown a live mirrored stream of themselves on the display, allowing them to monitor their gestures and make sure that they were in frame. Seed participant recordings were timed for 7 seconds and stopped automatically. Following the practice trial, seed participants were shown one (and only one) item from the meaning space as a single, orthographic word on screen, and followed the same procedure as the practice trial. The target meaning for each participant was selected at random, with the meaning removed from selection when two videos had been recorded for that meaning, giving a total of 48 seed videos. We refer to these sets of seed gestures throughout as generation 0.

### 2.1.4. Procedure: main experiment

The 50 participants recruited for the main experiment were organised in pairs comprising 5 generations in five transmission chains. Figure 2 illustrates the structure of a transmission chain in the experiment. Pairs of participants were taken through a two-part procedure in which they were trained on a set of gestures, and then asked to communicate with their partner, using only gesture. Participants were seated in individual experiment booths, and communication was enabled through video streaming

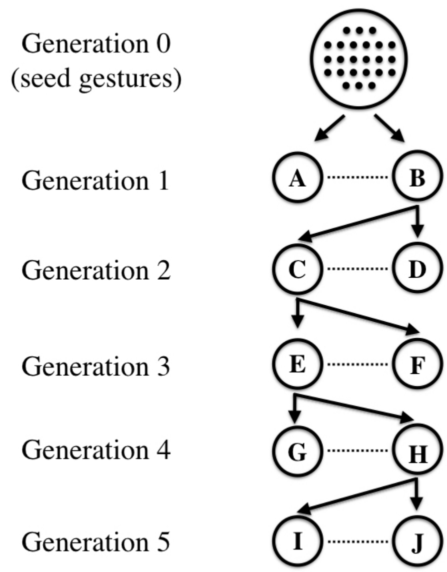



Figure 2: Transmission chain structure for experiment 1. Solid arrows represent transmission, dashed lines represent interaction. The seed gestures (generation 0) act as the starting sets for each chain. Participants in the first generation learn from the seed participants. They then communicate with each other and pass on their output (randomly selected from one of the two participants) as training for the second generation. This process repeats for five generations in total.

between two networked computers. As in the seed collection stage, participants were instructed not to use spoken language or manual spelling, and to remain seated throughout the task.

During training, participants were presented with videos of another person gesturing and asked to identify the correct meaning of the gesture. Each training trial consisted of the presentation of a gesture video, with simultaneous presentation of the meaning grid (example shown in figure 3). Participants could make their guess by clicking on a word from the grid at any point during the video, or at the end. Once a meaning had been selected, participants were given feedback: they were told whether they were correct or incorrect, and shown the correct interpretation of the gesture. The gesture video was then played again in full, without the opportunity to interrupt. Participants were subsequently asked to copy the gesture, and given a 3 second countdown to prepare themselves for recording. During recording, they were shown the mirrored live stream

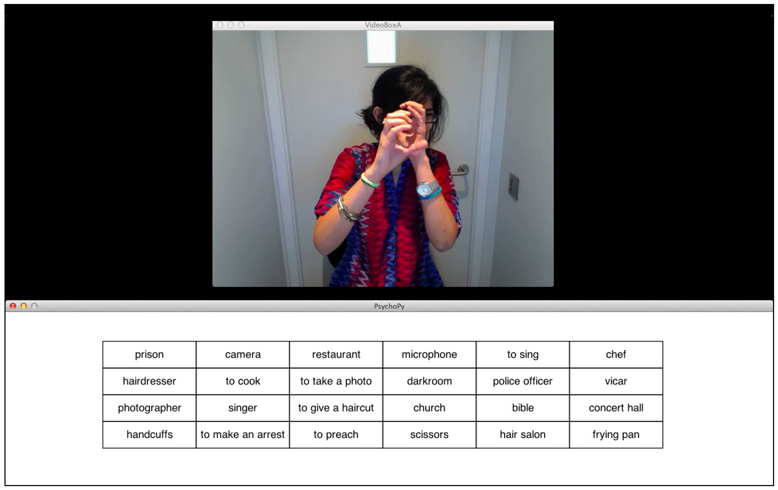



Figure 3: Screenshot example from a training trial, from a seed gesture for *camera*. The VideoBox window (top centre) presents pre-recorded videos in training as well as live stream during recording and testing. The Psychopy window (bottom) presents the meaning grid for interpretation, and shows instructions and feedback.

of themselves. The duration of recording was not pre-set. Participants were instructed to press the space bar to end recording. Each round of training consisted of 18 trials (18 out of the total 24 items in the meaning space).

The training items were selected randomly, and were balanced across the thematic and functional dimensions: 3 items from each theme were used and either 4 or 5 items from each functional type. The same 18 items were presented in each round of training, and were the same for both participants in a pair, though the order of presentation was randomised for each participant at each round. Participants completed 2 rounds of training (36 trials in total). All 24 items in the meaning space were presented in the meaning grid used for interpretation of the gestures. For each participant, the position of items in the grid was randomised, but remained constant for the duration of the experiment.

For participants in generation 1, a training set was generated by randomly selecting one of the two seed gestures from generation 0 for each meaning. For subsequent generations the training set consisted of gestures produced in the testing stage by a randomly selected participant from the pair in the previous generation. This meant that the full set of gestures from a single participant was transmitted and used as the model for the next generation. [1]

In the testing stage, participants took turns to communicate (as director) and interpret (as matcher) items in the meaning space, with each participant in a pair producing a gesture for each meaning once and interpreting a gesture for each meaning once (48 trials total). The order of presentation for target meanings was randomised.

As director, the participant was presented with an item from the meaning space. They were given 3 seconds with only the target item on screen, followed by a 3 second countdown to recording. The target meaning remained on screen throughout the trial. During recording, the participant performed their gesture, again seeing their own image mirrored onscreen, with a live, unmirrored stream sent to their partner's networked computer. The director was able to stop recording and turn off streaming by pressing the space bar, at which point they had to wait for their partner's interpretation to continue. The matcher could also stop streaming at any time by pressing space bar, cutting short the gesture. Gesture videos were therefore recorded from the beginning of the trial until the recordings was terminated by either director, or matcher (see below).

In the role of matcher, the participant had a short wait whilst the target item was presented to their partner, then a 3 second countdown to the video stream from their partner. Once streaming began, the same grid of meanings that appeared in the training stage was displayed. The matcher could make their guess (by clicking on an item in the grid) by first pressing the space bar to stop streaming, or by waiting for the director to stop recording. This ensured that timings reflected how long it took a matcher to comprehend a gesture, rather than the time it took them to find their responses in the grid of meanings. Following the matcher's response, both participants were given full feedback. They were shown whether the interpretation was correct or incorrect, and both the target item and the meaning selected by the matcher were presented on screen orthographically. Participants swapped roles after each trial, taking it in

[1] We transmitted the output from one participant only in order to simplify the learning process and prevent regularities from being obscured by potentially conflicting output from two participants. However, understanding learning from conflicting input is likely an interesting question in itself (Singleton & Newport, 2004; Smith et al., 2017b).

turns to be director and matcher for the duration of the study.

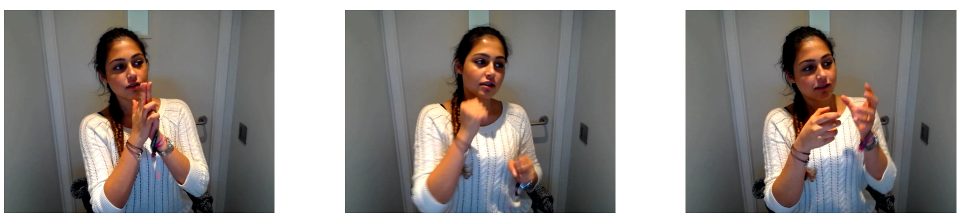
Throughout the testing stage, participants were shown a red timer on the right of the display, which ran during recording (and streaming) and accumulated across all trials. Participants were told that a monetary prize would be offered to the pair that was both fastest and most accurate, as calculated by overall time to complete the task minus a three second penalty for each incorrect interpretation.

### *2.2. Results: Qualitative data*

We first present a qualitative analysis of the gestures produced in the seed sets (generation 0) and the main experiment (generations 1-5). Gesture videos across all conditions in the experiment can be found in the University of Edinburgh's DataStore (http://dx.doi.org/10.7488/ds/2447; Motamedi et al., 2018).

#### *2.2.1. Seed gestures*

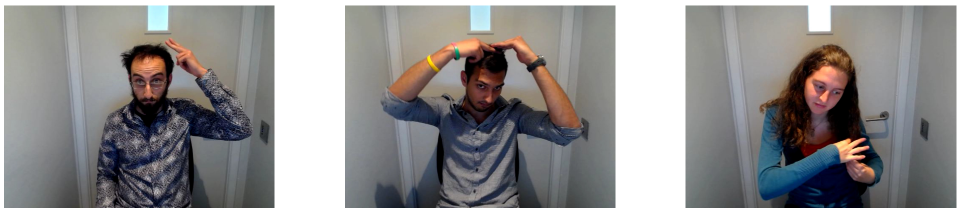
The gestures produced in the seed stage exhibit the global structure found in pantomime: large, highly iconic gestures which frequently depict elaborate scenes, such as that shown in figure 4, for the meaning *to make an arrest*.



(a) (b) (c)

Figure 4: Illustration of pantomime-like structure in seed gesture for *to make an arrest*. The participant illustrates a scene in which the presumed arresting officer draws and points a gun (a), before proceeding to run after the perpetrator (b), and then catching them (c).

Despite this, seed gestures sometimes lack features that could easily distinguish them from other meanings. For example, gestures produced for *hairdresser*, *hair salon* and *to give a haircut* (figure 5) all involve the gesturer miming cutting their own hair, but fail to clearly distinguish between the three meanings. This is largely as expected, since seed participants see only one meaning from the meaning space. In summary, these isolated gestures share a number of iconic features associated with their thematic category, but are highly inefficient and non-systematic.



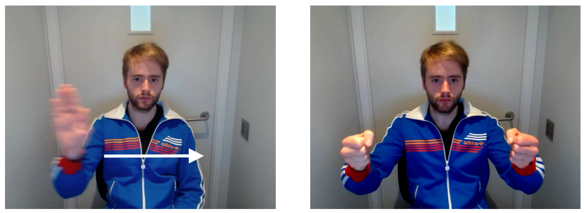
(a) (b) (c)

Figure 5: Ambiguity of gestures in the seed stage is exemplified by gestures for *hairdresser* (a), *hair salon* (b), and *to give a haircut* (c). All participants mime cutting their own hair, and there is little to distinguish between each meaning.

*2.2.2. Transmission and interaction stage*

From the seed set, over the course of five generations, gestures become more systematic, re-using and recombining gesture elements. For example, the gestures in figure 6 show productions for *prison* at generation 1 and generation 5 in one chain. At generation 1, both participants gesture shaking the bars of a prison cell. These gestures, like their seed precursor, are holistic and iconic, a pantomime of being behind bars. However, the generation 1 gestures are shorter in length than the seed model (both participants have dropped the palm movement from the seed gesture).

By generation 5 this meaning is gestured with a concise, two-part sign. Both participants produce a roof gesture followed by a wrist-grabbing gesture indicating handcuffs. The structure of these gestures is no longer holistic, but segmented. Furthermore, segmented gestures are widespread and systematic. Figure 7a gives examples of gestures from the same generation 5 participant shown in figure 6c, for the meanings *prison*, *church* and *hair salon*, all meanings from the *location* category. In each, the participant re-uses and recombines parts of signs; the roof sign is re-used as a category marker for location, followed by a thematic signal (in this case, a cuffing gesture for *prison*, a praying gesture for *church*, and a hair cutting gesture for *hair salon*). In this way, participants recombine meaningful gesture elements to systematically signal similarities and differences between meanings. Importantly, this set reflects not simply a single participant but a gradual development across chains.

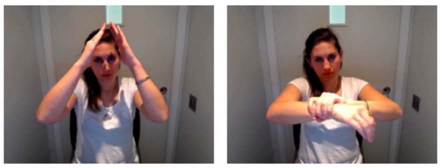


(a) Generation 0 gesture for *prison*.

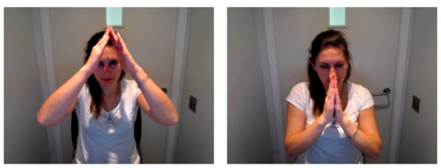
(b) Generation 1 gestures for *prison*. Gesture on left used as model for generation 2.

(c) Generation 5 gestures for *prison*.

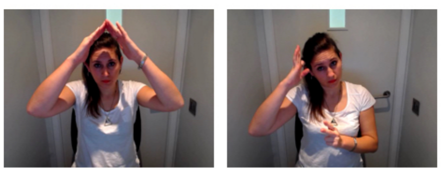
Figure 6: Over generations (a, b, c) in a chain, gestures for *prison* become more systematic across participants. The participant at generation 0 (seed video) moves a palm from right to left before shaking the bars of prison cell. Participants in generation 1 repeat the bar-shaking gesture, but drop the palm movement in the seed gesture. However, by generation 5, the gestures have changed. Both participants use a roof gesture, followed by a cuffing gesture to communicate the same meaning.



Gesture for *prison.*

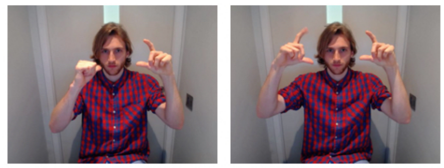


Gesture for *church.*

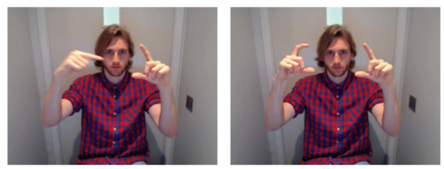


Gesture for *hair salon.*

(a) Re-use of the roof gesture across signals in the functional category of *location*.



Gesture for *photographer.*



Gesture for *camera.*

Gesture for *to take a photo.*

(b) Re-use of the camera gesture across signals in the thematic category of *photography*.

Figure 7: Segmentation into sequential, systematically re-used parts along the functional category (a), and the thematic category (b). Gestures in (a) communicate *prison*, *church*, and *hair salon*, and were produced by a participant in chain 3, generation 5, and show use of the same location marker, a roof shape. Gestures in (b) communicate *photographer*, *camera* and *to take a photo*, and were produced by a participant in chain 4, generation 5, and show use of the same thematic marker, a camera gesture.

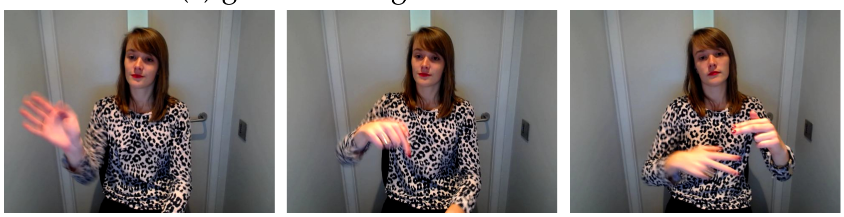
Gestures in later generations also demonstrate the re-use of thematic signals. Figure 7b illustrates a gesture for *photographer*, *camera* and *to take a photo* (thematic category of photography), in generation 5 of a chain. The camera shape is used as the sole signifier across all meanings for this category, paired with either a point-at-self for *photographer* or a point-at-object for *camera*, which signal the functional category (person or object).

Figure 8 illustrates how this process proceeds: a highly iconic pantomime is reanalysed as a symbolic grammatical marker. In generation 1, the gesture for *hairdresser* involves a pantomime in which the hairdresser waves to the customer, motions them to sit down, and mimes cutting hair. The gesturer finally points to herself, an indication of the person category. This hand wave gesture is repeated for the same meaning at generation 2, and by generation 3 has spread to other meanings within the same thematic dimension. By generation 5, the same element has been grammaticalised as a functional category marker, re-used throughout the functional category for *action*. This claim is supported by changes in form, including increasingly restrained movement and a decoupling from directed eye-gaze and facial expression; it is no longer an iconic representation of a greeting, but a systematic, symbolic marker. More generally, this example illustrates the finding that combinatorial systems emerge by generation 5 as a product of cumulative reanalysis.

### *2.3. Results: Quantitative results*

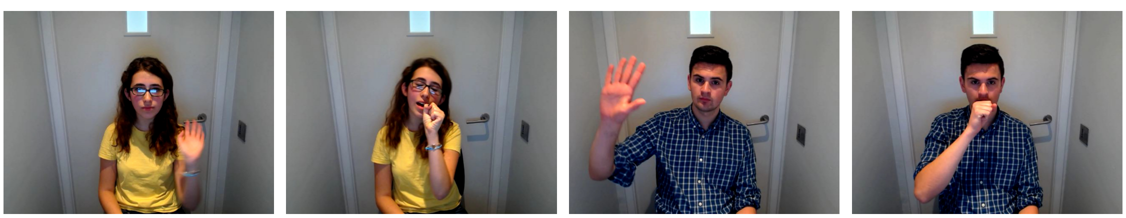
Recall that the presence of both interaction and transmission in this experiment was predicted to lead to gestures that are both communicatively efficient and systematic. Gibson et al. (Forthcoming) describe the main intuition behind efficiency as follows: an efficient language should enable a speaker to transmit many different messages successfully with minimal effort (p. 3). When measuring efficiency in an experimental context, different papers have operationalised different aspects of this intuition, depending partly on the properties of the experimental setting. For instance, Kirby et al. (2015) emphasise the *successful* transmission of information, by measuring the expressivity of the system. Other work (Fay & Ellison, 2013; Fay et al., 2013) focuses on *production effort*. Because our experimental setup shares features with that of the latter (i.e. it starts from a situation in which there are no conventions, and communication conventionalises over time), we will focus on production effort, measuring the efficiency of gestures using a combination of gesture length and

(a) generation 1 gesture for *hairdresser*



(b) generation 2 gesture for *hairdresser*

(c) generation 5 gestures for *to give a haircut*



(d) generation 5 gestures for *to sing*

Figure 8: Reanalysis of 'wave' pantomime as a functional category marker. The wave gesture starts out as an iconic depiction of a hairdresser's interaction with a customer (a). It is maintained in generation 2(b), and used for thematically related meanings in generations 3 and 4 (not shown). By generation 5 (c, d), the gesture has been reanalysed as a marker for the action category.

the number of repetitions within a gesture. We measure systematicity in two ways. First, we use entropy to measure the internal consistency of gestures for each participant. Higher entropy indicates use of a distinct set of idiosyncratic gestures to describe each meaning; low entropy indicates re-use of gestures from a limited set. We also use a measure of structure based on the presence of marking on the functional dimension (for the categories *person*, *location*, *object* and *action*). In terms of these measures then, we predict that over generations, (i) gestures will shorten in length and repetitions will be reduced, showing increased efficiency and (ii) gestures will decrease in entropy and involve more functional markers[2].

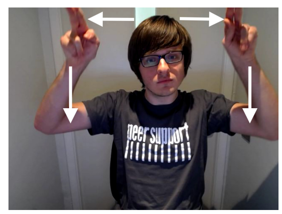
Gesture sequences for each meaning were coded by the first author for shape and handedness [3]. The shape parameter gave a characterising shape to each element in a gesture, such as *Thumb*, *Book*, or *Box*. Relevant elements such as direction could be added to distinguish, for example, between a point at an object (point-at-object) and a point at the gesturer's

[2]Additional results measuring matching accuracy, alignment between participants, and transmission success across experiment 1 and subsequent experiments can be found in the Supplementary Materials.

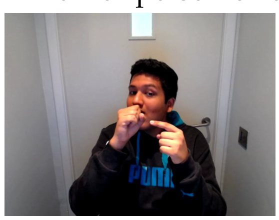
[3]A subset of gesture videos were coded by a second coder blind to the hypotheses of the experiment. The second coder analysed a sample of videos from one participant at each generation, randomly sampled from across the 5 chains, as well as a subset of the seed gestures that comprised at least one gesture from each functional category, and one gesture from each thematic category. This sample made up approximately 10% of all gestures. The second coder coded both the presence of functional marking, as well as coding gesture shapes in a sequence. The Cohen's kappa score representing inter-coder agreement for the presence of functional marking was 0.87, indicating very high agreement (Cohen, 1960). For the gesture shape and handedness coding, reliability does not rely on coders using the same tags as each other. Importantly, the tags themselves are not relevant, as long as the same tag is used to describe that gesture shape every time it appears. Therefore, what matters is how coders use the tags across gestures. We assessed reliability of gesture shapes following Sulik (2018). For all coded pairs of gestures for a given meaning, we calculated the Jaccard distance for each coder, which were averaged to find a mean Jaccard distance for each meaning (i.e. the extent to which the two gestures produced for that meaning were tagged in the same way by that coder). We then analysed the correlation between by-meaning Jaccard distance scores for each coder, to assess whether they distribute their tags in similar ways. We find a strong correlation for the use of gesture shape tags between the two coders ($r_s = 0.73, p < 0.001$). Coding guidelines and reliability analyses across all experiments can be found at https://osf.io/psxz6/

body (point-at-self). The handedness parameter coded whether each element in a gesture sequence was one- or two-handed. These parameters were combined for each element, creating an array for each gesture sequence. For example, [2-hand-Book, 1-hand-Point-at-object] would describe a gesture in which a participant gestures the shape of an open book with two flat palms and then points at the book object with a one-handed point. Gestures were also coded for the presence of marking along the functional dimension. Here a marker was defined as any part of signal meaningful to the entire functional category, and not just to the particular item in that category. Examples of typical markers are shown in figure 9.

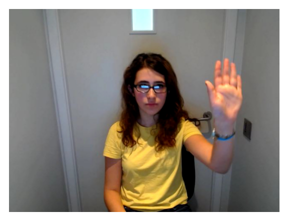
(a) marker for person category



(b) marker for location category



(c) marker for object category



(d) marker for action category

Figure 9: Examples of category markers used to distinguish items in different categories of the functional dimension of the meaning space. Examples are given from each category, for *person* (a), *location* (b), *object* (c) and *action* (d)

#### 2.3.1. *Efficiency*

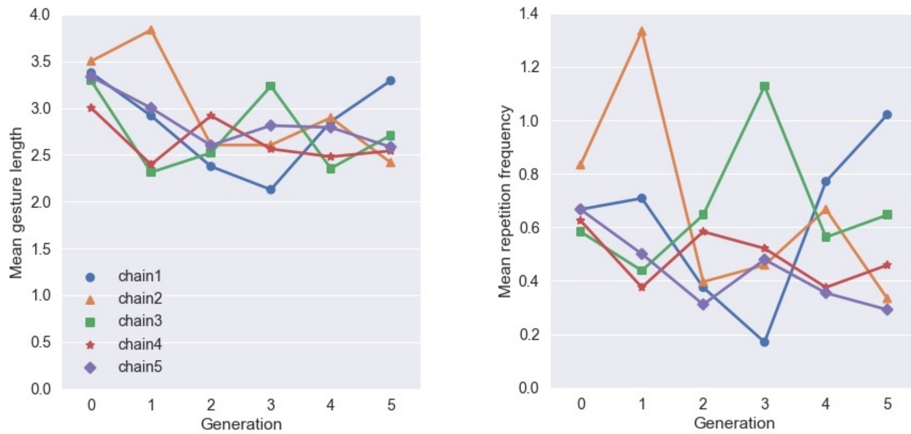
The most straightforward measure of efficiency is perhaps gesture length; all things being equal, shorter gestures encoding a given meaning are more economical. Gesture length, shown in figure 10a (left), was calculated as the number of individual elements coded for a particular meaning (e.g. the [2-hand-book, 1-hand-point-at-object] gesture described above would have a length of 2). We investigated change in the length of gestures over generations using a poisson mixed effects analysis, including generation as a fixed effect. Chain, participant and target meaning were included as random effects. We included random intercepts for all random effects, as well as a by-chain random slope of generation. The random effect of participant was nested within chain[4]. Model results for experiment 1 are shown in table 1. The model revealed a significant effect of generation, indicating that the length of gestures reduces as the systems are transmitted. However, figure 10a (left) suggests little change in gesture length following the seed generation. We ran a second model analysing the gesture length for generations 1 to 5 only; the results of the second model suggest that there is no change in gesture length over generations if the seed generation is not considered.

Visual inspection of gestures revealed that in some cases longer gestures were the result of repetitions. As repetitions are informationally redundant, they indicate a particular source of inefficiency in gestures. The frequency of repetitions within each participant's gestures is illustrated in figure 10a (right). Any individual gesture (as defined by the coding scheme) that was repeated within a sequence was counted as a repetition. For example, a sequence with a point-at-self, a mime of taking a photograph, then another point-at-self would have 1 repetition. Gesture elements which involve internally repeated movement (e.g., a camera gesture where the movement of pressing the shutter-release is repeated) were not counted as repetitions. The effect of generation on the proportion of repetitions was analysed using a poisson mixed effects model, including generation as a fixed effect. Analysis of the model did not reveal a significant effect of generation (see table 1). Participants do not show a systematic change in the use of repetitions over generations.

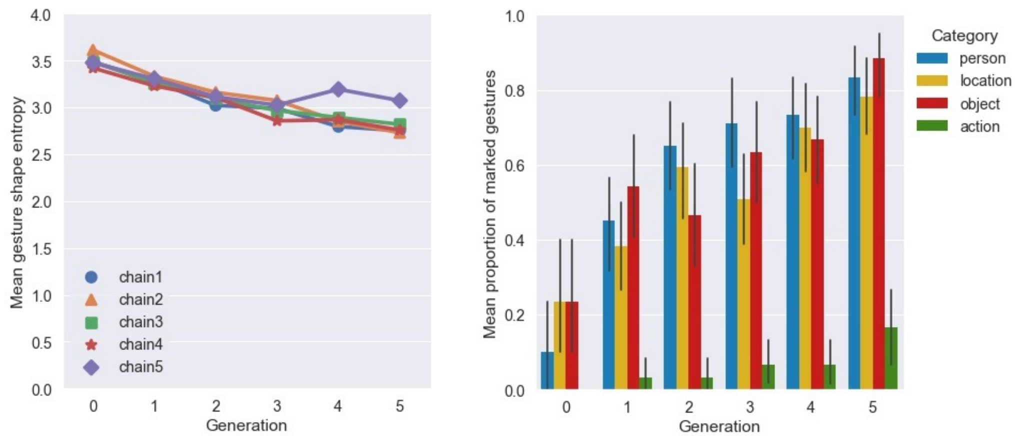
#### *2.3.2. Systematicity*

Recall that entropy is a measure of the consistency of the systems participants produced. For example, if a participant uses a pray gesture for *church*, and the same participant re-uses that gesture for *vicar*, *bible* and *to preach*, they show greater consistency (lower entropy) than if gestures

[4] All analyses here and henceforth were implemented using R (R Core Team, 2013) and *lme4* (Bates et al., 2015). All models were run using bound optimisation by quadratic approximation (*bobyqa*). Significance values for linear models were obtained using the R package LmerTest (Kuznetsova et al., 2017). For all analyses in this section, the same random effects structure is used, unless specified.




(a)




(b)

Figure 10: (a) Efficiency measures of mean gesture length (left) and mean number of repetitions in a gesture (right), shown for each chain (coloured lines with different markers). Gesture length is measured as the number of gesture shapes coded in a gesture sequence, repetitions as the number of repeated gesture shapes in a gesture sequence. Gestures show no systematic change in length or frequency of repetitions over generations. (b) Systematicity measures. Mean entropy of gesture shape (left) is shown for each chain (coloured lines with different markers). The proportion of functional marking (right) shows proportion of meanings in each category that contain any functional markers, and shown at each generation, with each coloured bar representing a corresponding functional category. Error bars represent bootstrapped 95% confidence intervals, across 5 transmission chains. Gestures become more systematic and consistent over the set a participant produces. Functional markers accumulate over generations in each category, though gestures for actions generally remain unmarked.

| | $\beta$ | $SE$ | $z$ | $p$ |
|---|---|---|---|---|
| **Gesture sequence length (all generations)** | | | | |
| generation | -0.03 | 0.01 | -2.27 | 0.02* |
| **Gesture sequence length (generations 1-5)** | | | | |
| generation | -0.01 | 0.02 | -0.46 | 0.64 |
| **Frequency of repetitions** | | | | |
| generation | -0.05 | 0.04 | -1.19 | 0.24 |
| **Functional marking** | | | | |
| generation | 0.71 | 0.12 | 6.48 | <0.001*** |
| category (verb or noun) | -3.51 | 0.92 | -3.81 | <0.001*** |
| generation * category | -0.19 | 0.18 | -1.08 | 0.28 |
| | $\beta$ | $SE$ | $t$ | $p$ |
| **Entropy** | | | | |
| generation | -0.12 | 0.02 | -7.10 | 0.002** |

Table 1: Model summary for measures in experiment 1. Outcome variables are given in bold type with fixed effect parameters underneath. For each fixed effect, we give the beta value, the standard error, the z-score or $t$ statistic (where appropriate) and the p-value.

for those meanings bore no resemblance to one another. We calculated the entropy of gesture sets used at each generation, based on the codes for individual gestures described above. Productions from a participant were pooled and entropy ($H$) was calculated over the atomic gestures ($x$) produced (i.e., not gesture sequences), given as

$$H = -\sum p(x) \log_2 p(x)$$

where entropy is summed over unique gestures in a set of gestures. For generation 0, the entropy was calculated over the seed sets used for that chain. Results are illustrated in figure 10b (left). We ran a linear mixed effects model predicting entropy from generation, including a random intercept for chain, and a random slope of generation. Participant and target meaning are not included in the random effects structure here as entropy is calculated over the set of gestures for a participant, giving one entropy value per participant. Our model revealed a significant effect of generation (table 1). As the systems are transmitted, the sets of gestures participants use become more consistent and less variable, with participants in later generations using fewer individual gestures in higher frequencies.[5]

Our second measure of systematicity is the use of functional markers. We counted the frequency of markers used in each generation of each chain, for each category in the functional dimension (*person*, *location*, *object* and *action*; see figure 10b (right)). Category types were collapsed into two categories: either nouns (*person, location, object*) or verbs (*action*), to examine the emergence of broad functional categories. We ran a logistic mixed effects model predicting the presence of functional marking from generation and category type (noun or verb, with noun as the baseline category), as well as their interaction. The model revealed a significant effect of generation and category type but no significant interaction (see table 1). These result suggest that marking for functional categories was introduced and increased as the systems were transmitted. Interestingly, marking introduced a distinction between noun and verb categories: the former tend to be marked, whilst the latter often re-

[5]This pattern is remarkably consistent across chains. The underlying distributions in each chain are different from each other, but still produce similar entropy values. See Supplementary Materials for more details.

main unmarked (though the presence of verb marking does increase over generations).

### 2.4. Experiment 1 summary

Experiment 1 demonstrates that systematic signals in a gestural communication system can emerge from largely unstructured pantomime over the course of five simulated generations in the lab. We do not find systematic evidence of a reduction in communicative efficiency— gesture length and a moderate amount of repetitions are maintained across generations. However, the accompanying increase in systematic structure reflects a reduction in the pool of gesture elements such that gestures from a more limited set are re-used to systematically distinguish items in the meaning space. By generation 5, gestures are no longer complex, idiosyncratic pantomimes, but comprise systems of segmented, interdependent signals. Marking systems emerge from early idiosyncratic gesture elements which evolve to comprehensively cover the meaning space by generation 5. This systematic structure facilitates learning by naive participants. These findings are consistent with previous work investigating the independent effects of transmission (Kirby et al., 2008; Smith & Wonnacott, 2010) on the emergence of systematicity in the lab. We also suggest that our results are consistent with studies that have found systematicity to emerge without transmission (Raviv et al., 2019; Winters et al., 2018; Nölle et al., 2018). In these cases, other manipulations (e.g. expanding meaning space, small training set) increase the pressure for simple, compressible systems in interaction. In the present case, when both interaction and transmission are present, systems with lower entropy and more structure emerge via competing pressures for learnability and communicative efficiency. As well as aligning with previous experimental research, this hypothesis offers an explanation for patterns found in natural languages generally (Kemp & Regier, 2012; Regier et al., 2015), and specifically in new sign languages (Goldin-Meadow, 2014; Senghas et al., 2004).

It is, perhaps, surprising that we do not see a clear indication of increased efficiency across generations, in light of previous work on interactional pressures (Fay & Ellison, 2013; Fay et al., 2010). Indeed, it is unclear what effect interaction is having in this case. Therefore, to demonstrate the effects of both interaction *and* transmission on the emergence of structure in the manual domain, experiment 2 tests what happens when each is isolated. Participants take part in either a transmission-only condition (where individual participants produce gestures without interacting, and those gestures are transmitted to a new learner) or an interaction-only condition (where two participants communicate repeatedly within a pair, without transmission to new learners). Following previous work (Kirby et al., 2008, 2015; Carr et al., 2016), we predict that transmission alone will lead to learnable, systematic signals that lack communicative efficiency. Conversely, without introducing any explicit pressure for simple, compressible systems, we predict that interaction alone will lead to shorter, more efficient signals that are nevertheless idiosyncratic and therefore less suitable for learning by naive users (Fay & Ellison, 2013; Theisen et al., 2010). In comparison, we predict that experiment 1 will represent the trade-off between the two pressures, with more systematicity in the gestures than those that develop through interaction alone, and more efficient systems than those that develop through transmission alone.

## 3. Experiment 2: isolating transmission and interaction

### *3.1. Methods: experiment 2*

#### *3.1.1. Participants*

35 participants were recruited from the same population as experiment 1 to take part in an interaction-only condition, and a transmission-only condition. Ten participants were recruited first for the interaction-only condition (aged 21 to 35, median age 22), followed by 25 participants for the transmission-only condition (aged 19 to 31, median age 23). Random assignment was not used, as the vastly different running times of the two conditions (approximately 90 minutes for the interaction-only condition and 45 minutes for the transmission-only condition) meant that participants were paid different amounts for participation and had to commit to experiments of different lengths. All participants were self-reported right-handed native English speakers, with no knowledge of sign language. Participants in the interaction-only condition were compensated £10 for participation, and participants in the transmission-only condition were compensated £5.

#### *3.1.2. Materials*

Materials were identical to those used in experiment 1.

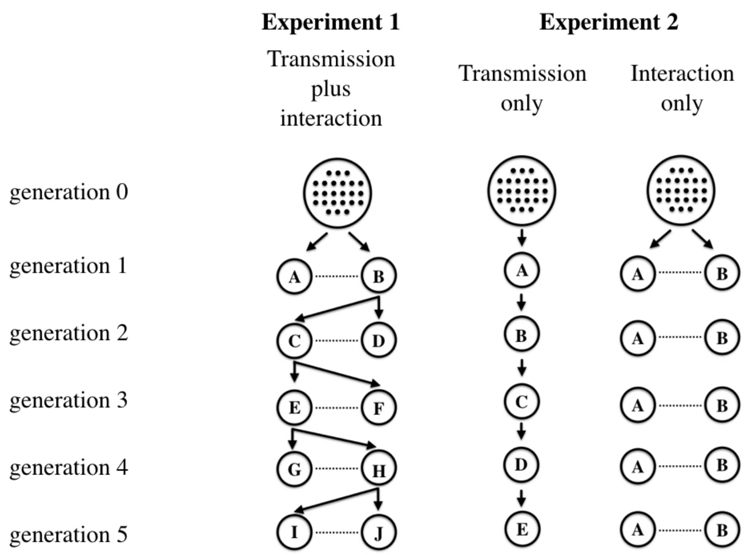



Figure 11: Transmission chain structure for all conditions in the two experiments. Solid arrows represent transmission, dashed lines represent interaction. The seed gestures (generation 0) act as the variable starting point for each chain or pair in each condition. Participants in the transmission-only condition only learn from previous participants and pass gestures on to new participants; they do not interact with other participants. Participants in the interaction-only condition only interact and do not pass on their gestures to new participants.

#### *3.1.3. Procedure*

The procedure for the two conditions was largely the same as experiment 1 (section 2.1.4). Participants in both conditions were instructed to use only gesture, not to speak or use manual spelling, and to remain seated throughout the task. While recording, participants were shown a mirrored image of themselves on-screen. Figure 11 illustrates the structure of each condition, compared with experiment 1.

*Transmission-only condition.* In the transmission-only condition, participants were trained on gesture sequences and then produced gestures for the same meanings. Both learning and production were done in isolation, with no communicative interaction. Participants were organised into 5 transmission chains of 5 generations, with a single participant at each generation (see figure 11).

As in experiment 1, during training, participants were shown videos of another person gesturing, from which they had to guess the meaning. Participants in generation 1 were exposed to a set of seed videos, while participants in generations 2-5 were shown a subset of videos recorded in testing by a participant from the previous generation in that chain. Participants were again shown gestures for 18 out of the 24 items in the meaning space, and completed two rounds of training.

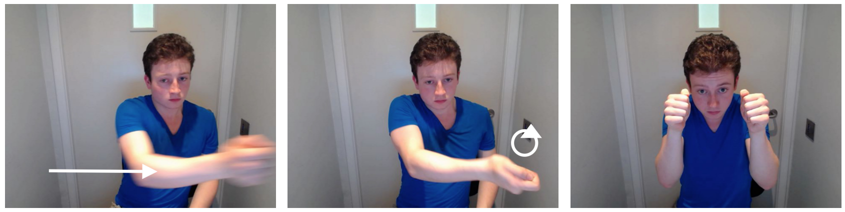
During testing, participants were presented with items from the meaning space and asked to communicate them using only gesture. After a 3 second countdown, recording started; once they had finished, participants could stop the recording by pressing space bar. Participants produced gestures for all 24 meanings in the meaning space. Unlike in experiment 1, participants in the transmission-only condition were not offered a bonus for fast and accurate responses, and were not explicitly timed; all pressures associated with communication were removed, ensuring that only the pressures associated with learning were present.

*Interaction-only condition.* In the interaction-only condition, pairs of participants repeatedly interacted with each other without transmission to new participants. Five pairs of participants took part in an initial training round, identical to experiment 1 (and the transmission-only condition) in terms of procedure. However, the training sets were always drawn from the set of seed gestures, and no training took place after generation 1.

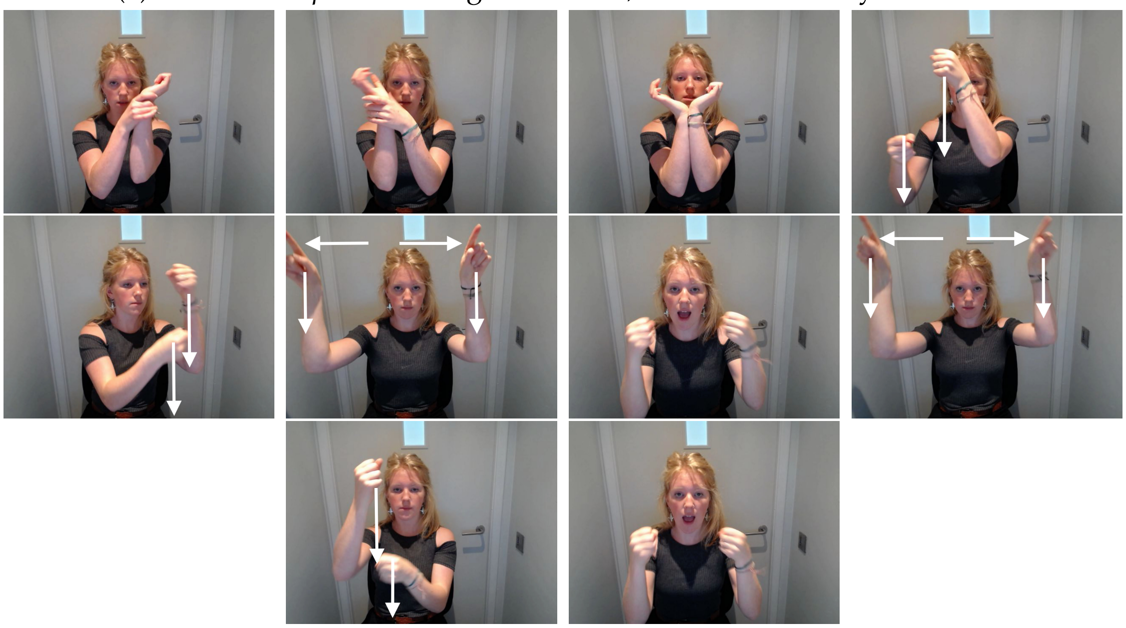
The testing stage in the interaction-only condition was identical to the testing in experiment 1. Participants had to communicate in pairs, taking it in turns to either communicate (as director) or interpret (as matcher) for all 24 items in the meaning space. Critically though, each pair took part in 5 consecutive testing rounds for the remainder of the experiment, parallel to the 5 generations of experiment 1 (see figure 11). As in experiment 1, participants were offered a bonus cash prize for the pair with the highest score, judged as a combination of speed and accurate interpretation of gestures; the pressure for communication was present in full, but the learnability pressure was reduced, since the systems were not transmitted to new generations.

### *3.2. Results: qualitative analysis*

We first look at typical examples from each condition in order to compare them qualitatively to the results from experiment 1.



(a) Gesture for *prison* from generation 1, transmission-only condition.



(b) Gesture for *prison* from generation 5, transmission-only condition.

(c) Gesture for *church* from a generation 5, transmission-only condition.

Figure 12: Functional marking and redundancy over generations in the transmission-only condition. Gestures are for the meaning *prison*, at generation 1 (a) and generation 5 (b) of the same chain, compared with the gesture for *church*, produced by the same generation 5 participant (c).

*Gestures in the transmission-only condition.* As in experiment 1, some use of marking emerges over generations in the transmission-only condition. Figure 12 shows gestures for *prison* for a generation 1 and 5 participant in the same chain. In generation 1, the gesture is holistic, lacking any structure that is systematically re-used across the set of gestures; by generation 5, systematicity emerges in the form of a location marker (the 'box' gesture). In contrast to experiment 1 however, the generation 5 gestures lack a clear two-part structure, and thematic dimension markers are rare. Further, redundancy (including repetition) appears to be common. This is shown in the same generation 5 gesture for *church*, shown in figure 12b. The participant consistently uses a location marker (though the roof-like properties are more noticeable here), which, along with the thematic cross and open-book gestures, are repeated. These features accord with our prediction that transmission results in some systematic structure, but not efficiency.

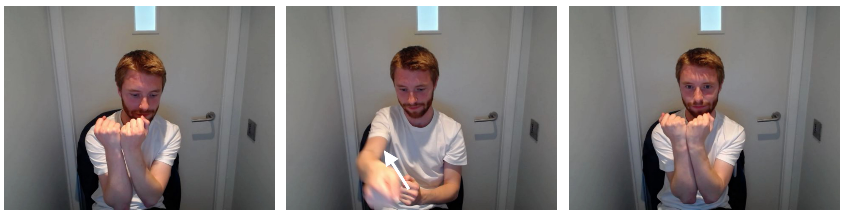
*Gestures in the interaction-only condition.* In the interaction-only condition, widespread systematic structure does not emerge. Figure 13 illustrates the development of gestures for *prison* from round 1 to round 5 for a single participant, in comparison with his gesture for *church* at round 5. While the gesture for *prison* is clearly shortened by round 5, the gestures have no clear segmentation and no marking that distinguishes the functional dimension (location). To summarize, in line with our predictions, gestures in the interaction-only condition show reduction in form, but lack the systematic structure found in experiment 1.

### *3.3. Results: Quantitative analysis*

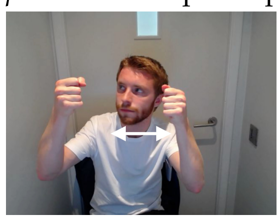
Here we present quantitative analyses comparing results from the transmission-only and interaction-only conditions to the results from experiment 1 (referred to as transmission+interaction) [6]. The coding scheme is identical to that described in section 2.3 [7].

[6] Additional analyses pertaining to communication accuracy, learnability and alignment can be found in the Supplementary Materials

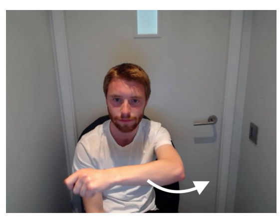
[7] Gestures were coded by a second coder, as described in experiment 1. Cohen's kappa for agreement on the presence of functional markers was 0.85, indicating very high agreement. Analysis of the gesture shape tags indicated a strong correlation between coders ($r_s = 0.83, p < 0.001$).



(a) Gesture for *prison* from a participant at round 1.



(b) Gesture for *prison* from a participant at round 5.



(c) Gesture for *church* from a participant at round 5.

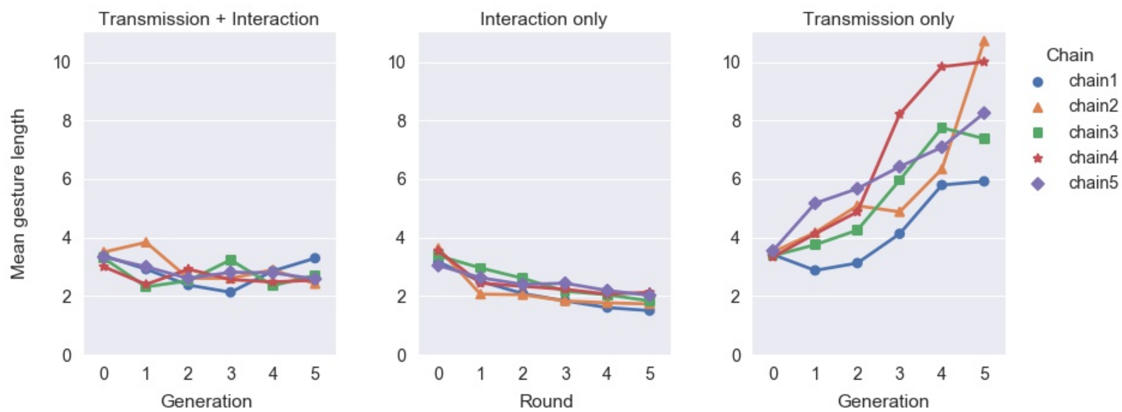
Figure 13: Lack of segmentation and marking of the meaning space dimensions across rounds in the interaction-only condition. Gestures shown are for *prison*, from round 1 (a) and 5 (b) for the same participant in the interaction-only condition, compared with a gesture from the same participant at round 5 (c), for the meaning *church*.

### *3.3.1. Efficiency*

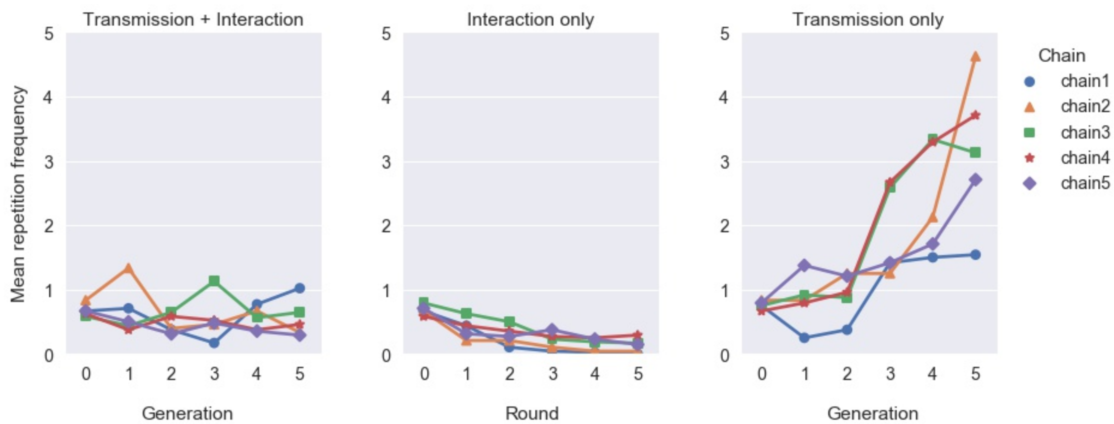
As before, efficiency is measured in terms of gesture length and frequency of repetitions. We ran a poisson mixed effects model predicting gesture length from generation (or round), condition and their interaction. Random intercepts for chain (or pair), participant (nested within chain or pair) and target meaning were also included, along with a by-chain random slope of generation [8]. Following the analysis of experiment 1, we report here the model for gesture length for generations 1-5 only (a model including all generations is given in the supplementary materials). The model revealed no significant change over generations for the baseline condition of transmission + interaction. However, we found significant interactions between generation and both the interaction-only condition and the transmission-only condition (table 2). When only interaction is present, gestures reduce in length over rounds, whereas when only transmission is present, gesture length *increases* (see figure 14a). Notably, gestures from the transmission + interaction condition represent a middle ground between the two, showing that similar gesture length is maintained across generations, in response to the trade-off between the two pressures.

Redundancy was measured as the number of repetitions within a single gesture sequence (as described in section 2.3). Figure 14b shows the mean number of repetitions within a single gesture, for each condition. We ran a poisson mixed effects analysis predicting number of repetitions from generation (or round), condition and their interaction. The model revealed no significant effect of generation, but a significant interaction between generation and each of the two other conditions (shown in table 2). This indicates that a larger decrease in redundancy is found in the interaction-only condition compared to the transmission+interaction condition. Conversely, redundancy *increases* in the transmission-only condition.

[8] All models use transmission+interaction (experiment 1) as the baseline condition. Note that chain is analogous to each pair in the interaction-only condition and generation is analogous to each round that a pair takes part in. The random effects structure described here is used in all subsequent analyses in this section, unless specified




(a)




(b)

Figure 14: Gesture length (a) and frequency of repetitions (b) for all conditions. Gestures become shorter and show fewer repetitions over rounds in the interaction-only condition, but show the reverse trend in the transmission-only condition. Gestures in the transmission + interaction condition show no change for length of repetitions over generations. Coloured lines with different markers represent each chain or pair.

| **Gesture sequence length (generation 1-5)** | $\beta$ | $SE$ | $z$ | $p$ |
|---|---|---|---|---|
| generation | -0.01 | 0.02 | -0.51 | 0.61 |
| condition - interaction only | -0.10 | 0.07 | -1.46 | 0.14 |
| condition - transmission only | 0.35 | 0.07 | 4.98 | <0.001*** |
| generation * condition - interaction only | -0.07 | 0.02 | -3.12 | 0.002** |
| generation * condition - transmission only | 0.20 | 0.03 | 7.17 | <0.001*** |

(a)

| **Frequency of repetitions** | $\beta$ | $SE$ | $z$ | $p$ |
|---|---|---|---|---|
| generation | -0.04 | 0.04 | 1.18 | 0.24 |
| condition - interaction only | -0.005 | 0.14 | -0.04 | 0.97 |
| condition - transmission only | 0.09 | 0.14 | 0.66 | 0.51 |
| generation * condition - interaction only | -0.29 | 0.05 | -5.52 | <0.001*** |
| generation * condition - transmission only | 0.35 | 0.05 | 6.78 | <0.001*** |

(b)

Table 2: Efficiency results from poisson mixed-effects regression models analysing the effect of generation and condition, as well as their interaction on (a) the length of gesture sequences, and (b) the frequency of repetitions in gesture sequences. Each table gives the outcome variable in bold, and each fixed effect underneath. For each fixed effect, we give the beta value, the standard error, the z-score, and the p-value.

*3.3.2. Systematicity*

Entropy of the gestures produced for each meaning was calculated for both the transmission-only and interaction-only conditions, using the same procedure described in section 2.3. Figure 15a shows the average entropy for all chains and pairs in each condition.

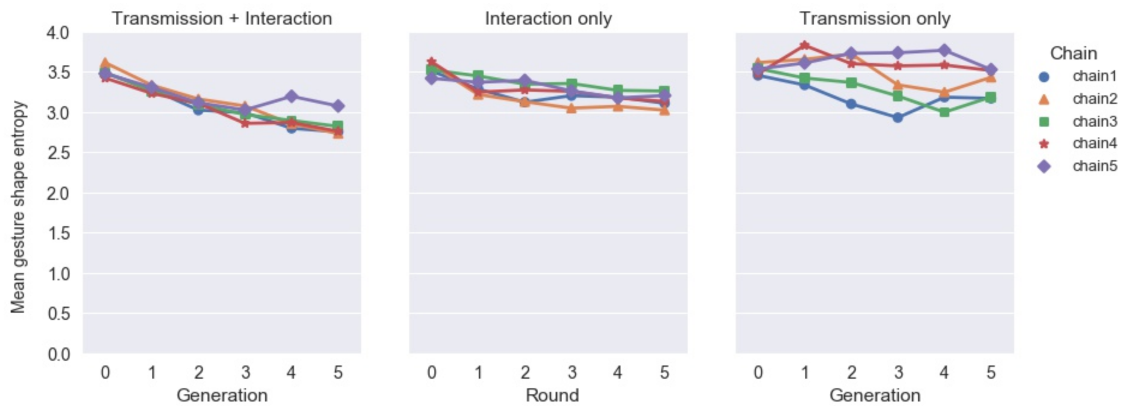
We ran a linear mixed effects model predicting entropy from generation, condition and their interaction, with a random intercept of chain and a random slope of generation. We also included a random intercept of participant, nested within chains. The model revealed a significant effect of generation for the transmission+interaction condition (baseline) and a significant interaction for both the transmission-only condition and the interaction-only condition. This indicates that entropy decreased over generations in the transmission+interaction condition, more so than in the interaction-only and transmission-only conditions. Further analysis on data from individual conditions revealed a reduction in entropy over generations for the interaction-only condition ($\beta = -0.06, SE = 0.01, t = -4.52, p = 0.04$) and a marginal trend for the transmission-only condition ($\beta = -0.05, SE = 0.02, t = -2.21, p = 0.08$). The combination of interaction and transmission thus lead to the greatest reduction in entropy, with only marginal evidence for a reduction in the other conditions.

We also examined the frequency of markers for categories in the functional dimension (*person*, *location*, *object* and *action*), as in experiment 1. This is shown for all conditions in figure 15b. We ran a logistic mixed effects model predicting marker frequency by generation, condition and their interaction. The model revealed a significant effect of generation, and a significant interaction between the generation and the interaction-only condition, but no significant interaction between generation and the transmission-only condition (table 3b). This indicates that category marking increased over generations in the transmission+interaction condition and transmission-only conditions but not in the interaction-only condition.

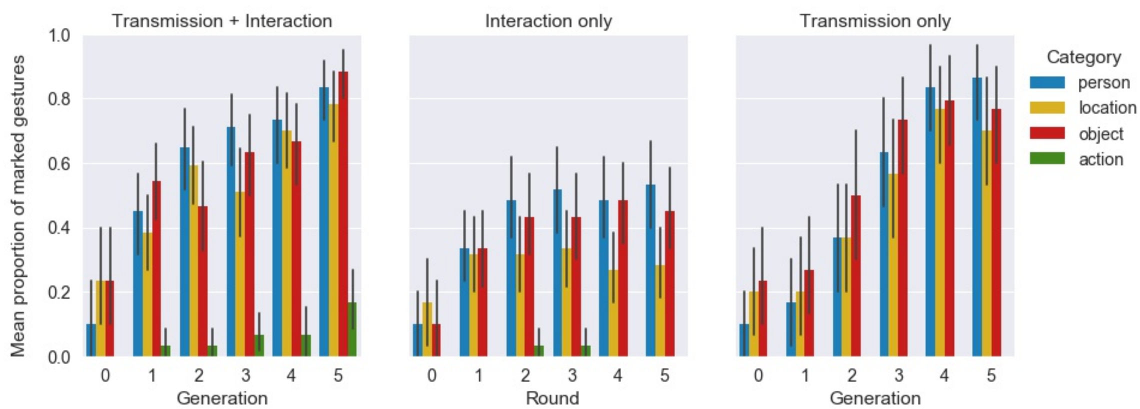
*3.4. Experiment 2 summary*

Experiment 2 investigated the gesture systems which emerged when interaction alone or transmission alone were present, in comparison to experiment 1, in which both were present.

Gestures from the interaction-only condition became shorter and showed fewer repetitions over rounds, while gestures in the transmission-only




(a)




(b)

Figure 15: (a) Mean entropy of gesture sets in each condition. Coloured lines with different markers represent mean values for each chain/pair. Entropy reduces to a greater extent in the transmission + interaction condition, than in the other two conditions. (b) Frequency of functional marking at each condition, shown at each generation for the categories of *person*, *location*, *object*, *action*. Error bars represent 95% bootstrapped confidence intervals. Strikingly, the frequency of functional markers increases only in conditions with transmission to new learners. Gestures in the interaction-only condition do not show the same cumulative increase in functional markers, compared to the other conditions.

| **Entropy** | $\beta$ | $SE$ | $t$ | $p$ |
| --- | --- | --- | --- | --- |
| generation | -0.12 | 0.01 | -10.83 | $<$0.001*** |
| condition - interaction only | 0.02 | 0.04 | 0.54 | 0.63 |
| condition - transmission only | 0.08 | 0.04 | 1.90 | 0.06 |
| generation * condition - interaction only | 0.07 | 0.01 | 6.01 | $<$0.001*** |
| generation * condition - transmission only | 0.1 | 0.02 | 6.53 | $<$0.001*** |

(a)

| **Functional marking** | $\beta$ | $SE$ | $z$ | $p$ |
| --- | --- | --- | --- | --- |
| generation | 0.75 | 0.09 | 8.30 | $<$0.001*** |
| condition - interaction only | -0.24 | 0.28 | -0.9 | 0.39 |
| condition - transmission only | -0.86 | 0.33 | -2.64 | 0.008** |
| generation * condition - interaction only | -0.54 | 0.09 | -5.83 | $<$0.001*** |
| generation * condition - transmission only | 0.14 | 0.11 | 1.25 | 0.21 |

(b)

Table 3: Systematicity results from mixed-effects regression models analysing the effect of generation and condition, as well as their interaction on (a) entropy of gesture sets, and (b) frequency of functional marker gestures. Each table gives the outcome variable in bold, and each fixed effect underneath. For each fixed effect, we give the beta value, the standard error, the z-score or $t$ statistic, where appropriate, and the p-value.

condition showed an increase in both length and repetitions. By contrast, gestures in the transmission + interaction condition demonstrate a trade-off between the two pressures, with gestures showing neither the reduction, nor the increase in length and repetition found in the other two conditions.

While gestures in the interaction-only condition demonstrate increased efficiency, they failed to show systematicity; gestures were not generally re-used across associated meanings. Instead, systematicity emerged in the transmission-only condition, as predicted if transmission introduces a pressure for systems to be more learnable by naive participants[9]. Overall, these results are compatible with the claim that competing pressures in transmission and interaction results in systems adapted for both learning and communication (Kirby et al., 2015). As such, only the combination of these pressures (here with a combination of transmission and interaction) will lead to language-like gestures that are efficient and demonstrate systematic recombination of segmented elements. However, as described above (section 2.1.4), our interaction procedure incorporates both a communicative task *and* explicit incentives for quick and accurate communication (a visual timer and a monetary prize). These explicit constraints could explain some of the differences between the two conditions that operationalise interaction and the transmission-only condition. In experiment 3, we re-run the transmission+interaction and the interaction-only conditions without these explicit incentives.

## 4. Experiment 3: efficient interaction without constraints

### *4.1. Methods and materials*

As with experiment 2, the two conditions were run separately rather than with random assignment to condition due to the different running times and payment rates of the two experiments. Apart from the removal of the explicit time pressure, the first condition replicated experiment 1 (transmission+interaction), and the second replicated the interaction-only condition of experiment 2.

[9]See Supplementary Materials for measures of learnability

#### 4.1.1. Participants

50 participants (aged 18-40, median age 22) were recruited from the University of Edinburgh careers website to take part in the transmission+interaction condition. The experiment took roughly 1 hour to complete and participants were compensated £7.50. Ten additional participants (aged 18-31, median age 21) were recruited for the interaction-only condition. Participants in the interaction-only condition took roughly 1.5 hours to complete the experiment and were paid £12 for participation. All participants were self-reported right-handed native English speakers with no knowledge of sign language.

#### 4.1.2. Materials

All materials used in the two conditions were the same as in experiment 1 (see section 2.1).

#### 4.1.3. Procedure

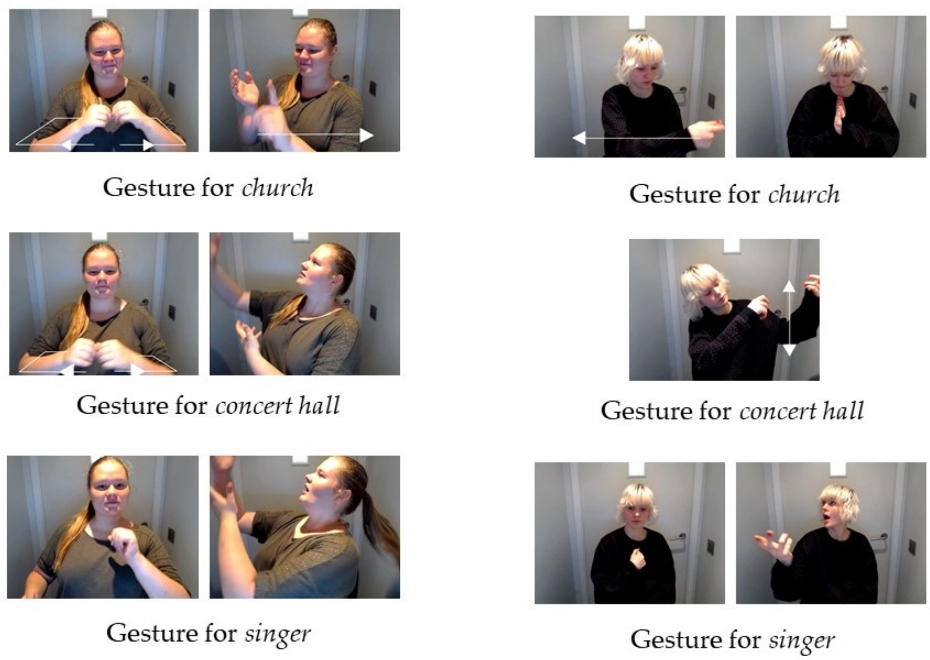
The procedure for both conditions was identical to that described in the relevant procedure sections with one exception: the explicit incentives for fast and accurate communication were removed. There was no timer, no prize, and no specific instructions as to how fast participants should be in communication.

### 4.2. Results: qualitative analysis

Gestures from experiment 3 (transmission + interaction and interaction-only) show similar qualitative structures to those in corresponding conditions in experiments 1 and 2. Gestures in the transmission + interaction condition, illustrated in figure 16a, show the re-use of gesture elements to signal similarities in meanings, both along the functional dimension and the thematic dimensions of the meaning space. Gestures in the interaction-only condition (shown in figure 16b) show some re-use of gesture parts, but tend to have much shorter gesture sequences that do not signal their shared associations.

### 4.3. Results: quantitative results

The analyses presented below contrast the two conditions of experiment 3 (transmission+interaction and interaction-only) with the transmission-




(a) Transmission + Interaction

(b) Interaction only

Figure 16: Examples of gestures from a) the transmission + interaction condition, and b) the interaction-only condition. Qualitatively, the gestures look similar to gestures from corresponding conditions in experiments 1 and 2. Gestures in a) demonstrate a two-part structure, signalling associations on both the thematic and functional dimensions of the meaning space. By contrast, gestures in b) do not systematically signal associations between meanings.

only condition of experiment 2 [10]. Gestures from experiment 3 were coded as described in section 2.3.[11]

#### 4.3.1. Efficiency

Gesture length and frequency of repetitions (figure 17) were both analysed using a poisson mixed effects model with generation, condition and their interaction as fixed effects (using the same random effects structure as described in section 3.3)[12]. As with previous results, we report here the gesture length model for generations 1-5. The gesture length model did not reveal a significant effect of generation (see table 4a), but demonstrated significant interactions between generation and condition, both for the interaction-only condition and the transmission-only condition. As in experiment 2, we find that gesture sequence length in the interaction-only decreases over generations, in comparison to the transmission + interaction condition, while it increases in the transmission-only condition.

The model analysing repetition frequency revealed similar results: a significant effect of generation as well as a significant interaction between generation and both other conditions (table 4b). Repetitions increase slightly in the transmission+interaction condition, though to a lesser extent than in the transmission-only condition, and decrease from generation 1 in the interaction-only condition. Again, without explicit pressure for efficient communication, redundancies in the conditions where interaction is present are still not as frequent as in the transmission-only condition. Overall, these results suggest that interaction by itself results in a pressure for increased efficiency.

#### 4.3.2. Systematicity

Entropy of participants' gesture sets and the frequency of gestures marking the functional dimensions of the meaning space are shown in

[10] As with the previous experiments, analyses pertaining to communicative accuracy, learnability and alignment can be found in the supplementary materials

[11] Gestures were coded by a second coder, as described in experiment 1. Cohen's kappa for agreement on the presence of functional markers was 0.82, indicating very high agreement. Analysis of the gesture shape tags indicated a strong correlation between coders ($r_s = 0.91, p < 0.001$).

[12] All models use transmission+interaction as the baseline condition, and the random effects structure described in section 3.3 is used in all subsequent analyses in this section, unless specified

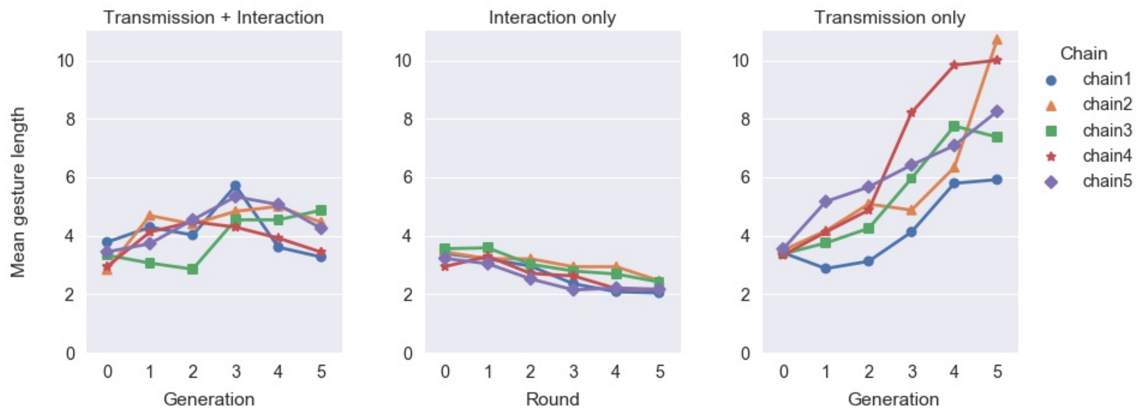
| **Gesture sequence length (generation 1-5)** | $\beta$ | $SE$ | $z$ | $p$ |
|---|---|---|---|---|
| generation | 0.01 | 0.02 | 0.67 | 0.50 |
| condition - interaction only | -0.26 | 0.07 | -3.93 | <0.001*** |
| condition - transmission only | -0.06 | 0.07 | -0.86 | 0.39 |
| generation * condition - interaction only | -0.11 | 0.02 | -5.04 | <0.001*** |
| generation * condition - transmission only | 0.18 | 0.03 | 6.23 | <0.001*** |

(a)

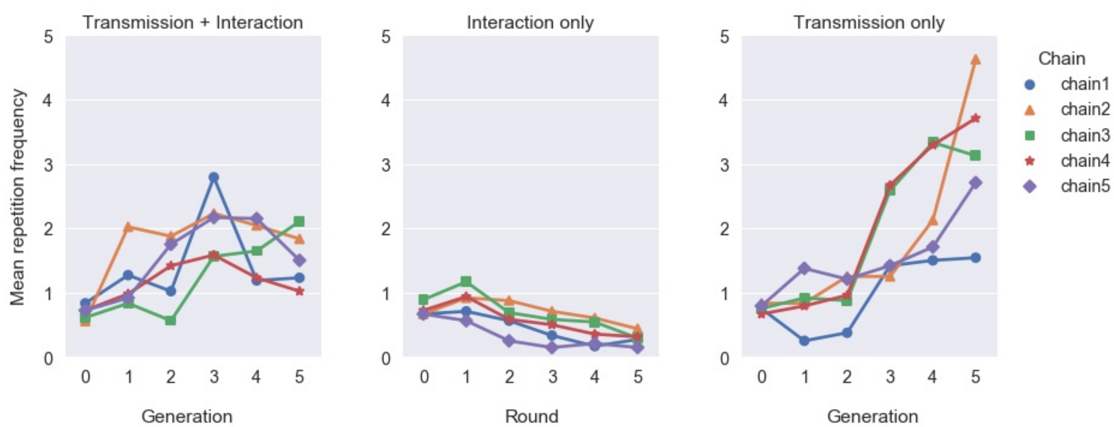
| **Frequency of repetitions** | $\beta$ | $SE$ | $z$ | $p$ |
|---|---|---|---|---|
| generation | 0.16 | 0.03 | 4.94 | <0.001*** |
| condition - interaction only | -0.02 | 0.13 | -0.18 | 0.86 |
| condition - transmission only | -0.24 | 0.13 | -1.79 | 0.07 |
| generation * condition - interaction only | -0.42 | 0.04 | -10.09 | <0.001*** |
| generation * condition - transmission only | 0.13 | 0.05 | 2.84 | 0.004** |

(b)

Table 4: Efficiency results from poisson mixed-effects regression models analysing the effect of generation and condition, as well as their interaction on (a) the length of gesture sequences, and (b) the frequency of repetitions in gesture sequences. Each table gives the outcome variable in bold, and each fixed effect underneath. For each fixed effect, we give the beta value, the standard error, the z-score, and the p-value.




(a)




(b)

Figure 17: Gesture length (a) and frequency of repetitions (b) for all conditions. Gesture sequences in the interaction-only condition become shorter over generations and use fewer repetitions, while gestures in the transmission-only condition become longer with more repetitions. Gestures in the transmission + interaction condition show no significant change in gesture length or repetitions over generations. Coloured lines and different markers represent the mean values for each chain/pair.

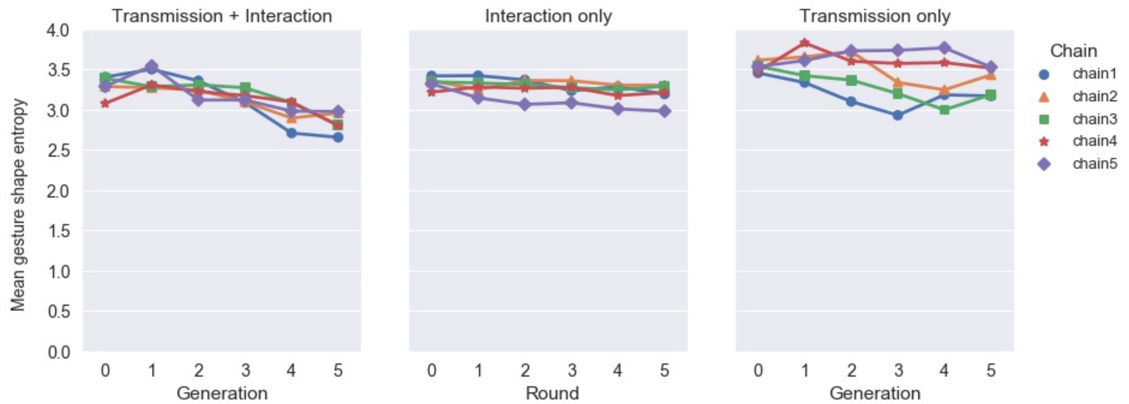
figure 18.

*Entropy.* A linear mixed effects regression model predicted entropy from generation, condition and their interaction, with random intercepts of chain and participant. We also included a by-chain slope of generation, and the random effects for participant were nested in chains. The model revealed a significant effect of generation for the transmission+interaction condition (baseline, table 5a). The model also revealed a significant interaction for both other conditions. These results replicate the results from experiment 2; entropy reduces over generations in the transmission+interaction condition, and this same reduction is not found to the same extent in the other two conditions.

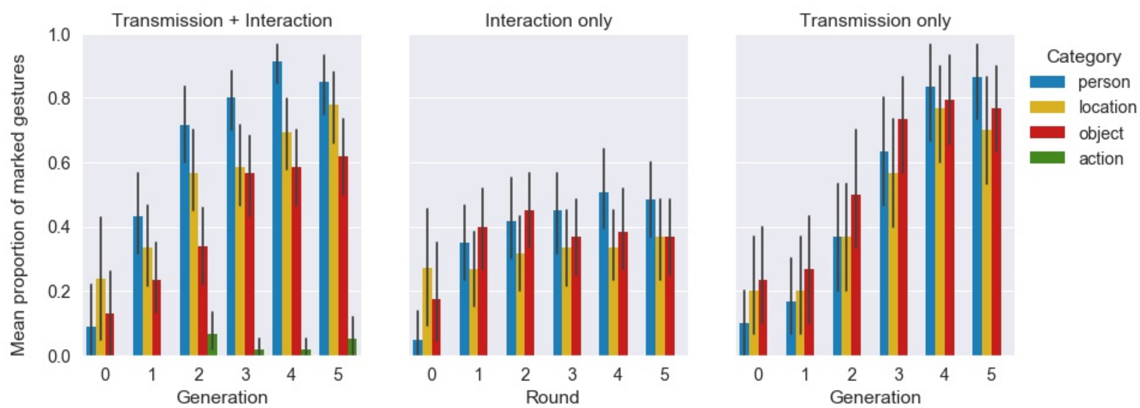
*Functional marking.* A logistic mixed effects model predicting frequency of functional marking by condition, generation and their interaction revealed a significant effect of generation for the transmission+interaction condition (table 5b). There was no significant interaction for the transmission-only condition. However, this interaction was significant in the interaction-only condition. These results replicate our previous findings: there is an increase in marking over generations for both the transmission-only and the transmission+interaction conditions, but not the interaction-only condition.

### *4.4. Experiment 3 summary*

In experiment 3, we removed the explicit incentives for quick and accurate communication (a visual timer and a monetary prize) from the two interaction conditions in order to rule this out as a potential explanation for the differences between these conditions and the transmission-only condition. We replicated our previous findings regarding systematicity (entropy) and functional marker frequency; even without these incentives, both transmission conditions differed from the interaction-only condition. Furthermore, we find similar results with regards to gesture length; gestures in the interaction-only condition show a reduction in length, while gesture length does not change consistently across generations 1-5 in the transmission + interaction condition. We do however, find that gestures in the transmission + interaction condition show an increase in the use of repetitions over generations, in contrast to the results from experiments 1 and 2. Given that this does not lead to a comparative increase in length,




(a)




(b)

Figure 18: (a) Mean entropy of gesture sets in each condition. Coloured lines with different markers represent means for each chain/pair. Entropy reduces across generations in the transmission + interaction condition, but does not reduce to the same extent in the other two conditions. (b) Frequency of functional marking at each condition, shown at each generation for the categories of *person*, *location*, *object*, *action*. Error bars represent 95% bootstrapped confidence intervals. The frequency of functional markers only increases in conditions with transmission to new learners.

| **Entropy** | $\beta$ | $SE$ | $t$ | $p$ |
| --- | --- | --- | --- | --- |
| generation | 0.27 | 0.04 | -6.90 | <0.001*** |
| condition - interaction only | -0.02 | 0.04 | -0.48 | 0.63 |
| condition - transmission only | 0.19 | 0.04 | 4.71 | <0.001*** |
| generation * condition - interaction only | 0.07 | 0.01 | 5.08 | <0.001*** |
| generation * condition - transmission only | 0.03 | 0.02 | 2.09 | 0.04* |

(a)

| **Functional marking** | $\beta$ | $SE$ | $z$ | $p$ |
| --- | --- | --- | --- | --- |
| generation | 0.80 | 0.10 | 7.76 | <0.001*** |
| condition - interaction only | 0.25 | 0.36 | 0.70 | 0.48 |
| condition - transmission only | -0.42 | 0.39 | -1.08 | 0.28 |
| generation * condition - interaction only | -0.66 | 0.10 | -6.14 | <0.001*** |
| generation * condition - transmission only | 0.09 | 0.14 | 0.66 | 0.51 |

(b)

Table 5: Systematicity results from mixed-effects regression models analysing the effect of generation and condition, as well as their interaction on (a) entropy of gesture sets, and (b) frequency of functional marker gestures. Each table gives the outcome variable in bold, and each fixed effect underneath. For each fixed effect, we give the beta value, the standard error, the z-score or $t$ statistic, where appropriate, and the p-value.

and in light of the reduced entropy, it is possible that these repetitions are part of the structural reorganisation of gesture sequences.

On the whole, we find that gestures in the transmission + interaction condition still represent a middle ground between the two pressures for measures of efficiency. This suggests that a communicative task in the absence of explicit incentives for speed still introduces a substantial pressure for efficiency. Moreover, the differences between the three conditions again suggest that when both transmission and interaction are present, there is a trade-off between them; gestures in the transmission+interaction condition are intermediate between the two isolated conditions. Importantly, we once again demonstrate that efficient, systematic structure emerges when both transmission and interaction are at play.

## 5. General discussion

The present study aimed to investigate the independent and combined contributions of interaction (using a system to communicate) and transmission (learning by new generations) in shaping emerging manual communication systems. Specifically, we were interested in how these processes facilitate the continuous evolution of pantomimes into language-like gestures, involving a shift from holistically structured, inefficient gestures to ones exhibiting systematicity and efficiency.

In experiment 1, we showed a combination of interaction and transmission leads to the emergence of language-like systems from pantomime. Systematicity increased through the introduction and expansion of interdependent re-usable signals conveying the dimensions of the meaning space. However, gesture length and repetition frequency showed no consistent changes over generations, which did not clearly suggest any change in the efficiency of gestures participants produced.

In experiment 2, we tested the effects of each mechanism in isolation. We found that, although gestures in each condition exhibited some increase in the language-like properties we measured, they did not do so to the same degree as in experiment 1. As predicted, without a pressure for learnability, gestures in the interaction-only condition showed increased efficiency, and consistency but gestures remained idiosyncratic and no widespread use of functional category marking developed. This extends the results of previous studies to the gestural modality (Garrod et al., 2007; Theisen et al., 2010; Fay et al., 2010; Fay & Ellison, 2013; Kirby et al., 2015). Without new learners being introduced, there is little pressure for learnability; gestures used within a pair do not need to be predictable across the meaning space for communicative success. Signals in these cases lack systematic structure, and individual signals remain relatively independent of each other.

By contrast, the systems that emerge in the transmission-only condition showed widespread use of marking to signal the functional dimension, comparable with experiment 1. They demonstrated segmentation and increased systematic structure over generations. However, gestures were relatively less efficient. This again extends previous findings which demonstrate a cumulative increase in structure as signals are transmitted to new learners (Beckner et al., 2017; Kirby et al., 2008; Verhoef et al., 2014). While a pressure for learnability led to an interdependent system of re-used and recombined signals (Kirby et al., 2015; Raviv et al., 2019; Nölle et al., 2018), without the pressure for efficient communication, gestures produced by individuals in the transmission-only condition were much longer and exhibited large-scale redundancies through repetition. Notably, it is unclear why participants in the transmission-only condition would produce such long and repetitive sequences, given that they are somewhat effortful to produce. Though it is possible that this is an artefact of the task (i.e. sitting in front of a camera may carry some implicit expectations), similar results have been found in a task where signals were created with buzzer presses (Kempe et al., 2017). The motivations for these inefficient gesture sequences are unclear in this case, but may be an interesting question in itself. Whatever these motivations may be, however, our results clearly demonstrate that the use of gestures in interaction moderates the production of long, redundant gestures. In experiment 2, the lack of change in efficiency measures in the transmission + interaction condition demonstrates the trade-off between transmission and interaction. Gesture length and repetitions neither increase (as with transmission) or decrease (as with interaction), but represent an intermediate development.

Finally, in experiment 3 we removed the incentives for quick and accurate communication from the experimental design in the transmission+interaction and interaction-only condition. Because these were present in addition to the communicative task in experiments 1 and 2, removing them allowed us to test the extent to which they were driving some of the differences between our conditions. As expected, our findings with respect to systematic structure were replicated. Furthermore, explicit incentives were not necessary for increased efficiency over generations in the interaction-only condition; as in experiment 2, reductions in gesture sequence length and repetition frequency were both found. In the transmission+interaction condition, however, instead of a decrease we found a slight increase in the number of repetitions, though this was not reflected in increased gesture length. Importantly though, the transmission+interaction condition clearly exhibits the signature effects of both mechanisms. Efficiency is lower than in the interaction-only condition, but still higher than in the transmission-only condition. The implicit pressure for communicative efficiency is therefore still at work.

The set of experiments described here demonstrate the effects of transmission and interaction on an evolving manual linguistic system. Our findings support the hypothesis that—independent of modality—pressures for learnable and communicatively effective systems drive the emergence of language-like structure. Crucially, both pressures must be present for the emergence of signals that are *both* systematic and efficient (Regier et al., 2015; Theisen-White et al., 2011). Specifically, neither transmission alone nor interaction alone led to language-like structures in our experiments. Only when both mechanisms worked together did we see the emergence of gestures that maintained communicative efficiency whilst at the same time developing structure through the systematic recombination of segmented signals. The miniature artificial sign languages that evolve in these experiments show evidence of adapting to the specific pressures at play in each condition; systems in the interaction-only condition become suited to efficient communication within a pair, whilst systems in the transmission-only condition lack communicative efficiency, but demonstrate systematic structure which signals the dimensions of the meaning space. These results are corroborated by naturalistic data from emerging sign systems. Homesigns, used by individuals, lack regularity and exhibit low rates of conventionalisation (Richie et al., 2014; Goldin-Meadow et al., 2014). Emerging sign languages in their earliest stages begin to show stabilisation and conventionalisation, but they lack the relative consistency of older sign languages; as the systems are used in interaction and transmitted to new learners, the languages further stabilise and begin to develop consistent and regular structures across signers (Sandler et al., 2005; Padden et al., 2010; Goldin-Meadow et al., 2014). Our results reveal similar patterns; gestures in early generations (generations 0 and 1) lack regularity and show little evidence of conventionalisation. Gestures that are used between pairs of participants without transmission become more conventionalised and efficient, but lack systematic structure. But through the repeated use *and* transmission of the systems, gestures become systematic and regular within a chain. In particular, our results show the development of categorical markers that distinguish between nouns and verbs in the meaning space, consistent with early development of such categories in emerging sign languages (Tkachman & Sandler, 2013; Padden et al., 2013; Goldin-Meadow et al., 2014). We find that segmentation and grammaticalisation of holistic gestures can occur following transmission, in response to pressures for learnable, compressible systems, aligning with research on segmenation in NSL (Senghas et al., 2004). Furthermore, our results exhibit a pattern consistent with the results of Goldin-Meadow et al. (2014), where stability of categorical distinctions increased over cohorts of NSL from Nicaraguan homesigners (who showed little stability) to second-cohort signers.

By using the manual modality, our experiments also potentially minimise the interference that prior linguistic knowledge may have had in previous research using artificial language learning experiments. More generally, this method allows the investigation of modality-independent mechanisms that affect language, but may also offer a platform for investigating modality-specific phenomena involved in the emergence of linguistic structure, such as the presence of iconicity in a system as it develops structure (Micklos, 2017).

## 6. Conclusion

Previous work on the naturalistic emergence of manual communication systems has shown how interaction between speakers in a community together with the introduction of new language learners leads to distinct linguistic features. At the same time, experimental research on the evolution of spoken and written languages in the lab has provided confirmatory evidence that both interaction and transmission are crucial to the emergence of structured, efficient systems. Here we combine two well-known paradigms—silent gesture and iterated learning—to investigate how linguistic structure emerges in artificial manual sign systems. By incorporating both interaction within generations and transmission of the system to naive learners in our experiments, we have shown that the combined effects of these mechanisms drive the gradual emergence of systematic and efficient gestures. When both are present, gestures showed an increase in conventionalisation, signal re-use, and functional marking, and a reduction in the redundancy that is characteristic of pantomime. The studies presented here extend previous experimental work, while potentially reducing the effect of prior linguistic experience on participants' behaviour. At the same time, they offer a parallel to the observation of structural development in natural sign languages. By providing an experimental complement to data from natural languages, this work illustrates how we can begin to bridge the gap between longitudinal, naturally occurring data on language emergence, with controlled hypothesis testing through experimentation.

## 7. Acknowledgements

We would like to thank Fiona Kirton, who collected data for experiment 3, and Hannah Fricke, who second coded the data for all three experiments. YM was funded by a Carnegie-Caledonian Scholarship from the Carnegie Trust for the Universities of Scotland. MS was funded by a British Academy Postdoctoral Fellowship (award number pf130034).

## 8. References

Aronoff, M., Meir, I., & Sandler, W. (2005). The Paradox of Sign Language Morphology. *Language*, *81*, 301–344.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48.

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, *2*, 1–16.

Caldwell, C. A., & Smith, K. (2012). Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PloS one*, *7*, e43807.

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2016). The Cultural Evolution of Structured Languages in an Open-Ended, Continuous World. *Cognitive Science*, *41*, 892–923.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cornish, H., Smith, K., & Kirby, S. (2013). Systems from Sequences: an Iterated Learning Account of the Emergence of Systematic Structure in a Non-Linguistic Task. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the Cognitive Science Society* (pp. 340–345). Austin, Texas: Cognitive Science Society.

De Vos, C. (2014). The Kata Kolok Pointing System: Morphemization and Syntactic Integration. *Topics in Cognitive Science*, *7*, 150–68.

Fay, N., Arbib, M., & Garrod, S. (2013). How to bootstrap a human communication system. *Cognitive Science*, *37*, 1356–67.

Fay, N., & Ellison, T. M. (2013). The cultural evolution of human communication systems in different sized populations: usability trumps learnability. *PloS one*, *8*, e71781.

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*, 351–86.

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

Gibson, E., Futrell, R., Piantadosi, S., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. P. (Forthcoming). How Efficiency Shapes Human Language. *Topics in Cognitive Science*, .

Giudice, A. D. (2012). The emergence of duality of patterning through iterated learning: Precursors to phonology in a visual lexicon. *Language and Cognition*, *4*, 381–418.

Goldin-Meadow, S. (2003). *The Resilience of Language*. New York, NY: Taylor and Francis.

Goldin-Meadow, S. (2014). The Impact of Time on Predicate Forms in the Manual Modality: Signers, Homesigners, and Silent Gesturers. *Topics in Cognitive Science*, (pp. 1–16).

Goldin-Meadow, S., Brentari, D., Coppola, M., Horton, L., & Senghas, A. (2014). Watching language grow in the manual modality: Nominals, predicates, and handshapes. *Cognition*, *136*, 381–395.

Goldin-Meadow, S., Butcher, C., Mylander, C., & Dodge, M. (1994). Nouns and Verbs in a Self-Styled Gesture System: What's in a name? *Cognitive Psychology*, *27*, 259–319.

Goldin-Meadow, S., & Feldman, H. (1977). The development of language-like communication without a language model. *Science*, *197*, 401–403.

Goldin-Meadow, S., & Mylander, C. (1983). Gestural communication in deaf children: noneffect of parental input on language development. *Science (New York, N.Y.)*, *221*, 372–374.

Goldin-Meadow, S., & Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. *Nature*, *391*, 279–281.

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*, 195–262.

Goldin-Meadow, S., So, W. C., Ozyürek, A., & Mylander, C. (2008). The natural order of events: how speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 9163–8.

Hall, M. L., Mayberry, R. I., & Ferreira, V. S. (2013). Cognitive constraints on constituent order: Evidence from elicited pantomime. *Cognition*, *129*, 1–17.

Haviland, J. B. (2013). The emerging grammar of nouns in a first generation sign language: Specification, iconicity, and syntax. *Gesture*, *13*, 309–353.

Healey, P. G. T., Swoboda, N., Umata, I., & King, J. (2007). Graphical language games: interactional constraints on representational form. *Cognitive Science*, *31*, 285–309.

Kegl, J., Senghas, A., & Coppola, M. (1999). Creation through contact: Sign language emergence and sign language change in Nicaragua. In M. DeGraff (Ed.), *Language Creation and Language Change: Creolization, Diachrony and Development* (pp. 179–237). Cambridge, MA: MIT Press.

Kemp, C., & Regier, T. (2012). Kinship Categories Across Languages Reflect General Communicative Principles. *Science*, *336*, 1049–1054.

Kempe, V., Gauvrit, N., & Tamariz, M. (2017). Iterated Teaching Can Optimise Language Functionality. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 651–656). Austin, TX: Cognitive Science Society.

Kirby, S. (2016). VideoBox. University of Edinburgh. Retrieved from http://edin.ac/2haREUz.

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 of the United States of America*, *105*, 10681–6.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology*, *28C*, 108–114.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and Communication in the Cultural Evolution of Linguistic Structure linguistic structure. *Cognition*, *141*, 87–102.

Klima, E. S., & Bellugi, U. (1979). *The Signs of Language*. Cambridge, MA: Harvard University Press.

Kocab, A., Pyers, J., & Senghas, A. (2014). Referential Shift In Nicaraguan Sign Language: A Transition From Lexical To Spatial Devices. *Frontiers in Psychology*, .

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26.

McNeill, D. (2000). Introduction. In D. McNeill (Ed.), *Language and Gesture* (pp. 1–10). Cambridge: Cambridge University Press.

Micklos, A. (2017). Iconic strategies in silent gesture: Perceiving the distinction between nouns and verbs. In *11th International Symposium on Iconicity in Language and Literature* (pp. 26–27).

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2018). Evolving artificial sign languages in the lab: from improvised gesture to systematic sign [dataset]. University of Edinburgh DataShare. Retrieved from https://datashare.is.ed.ac.uk/handle/10283/3191.

Nyst, V. (2010). Sign languages in West Africa. *Sign languages*, (pp. 405–432).

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

Padden, C., Meir, I., Aronoff, M., & Sandler, W. (2010). The grammar of space in two new sign languages. *Sign Languages: A Cambridge Language Survey*, (pp. 570–592).

Padden, C., Meir, I., Hwang, S.-O., Lepic, R., Seegers, S., & Sampson, T. (2013). Patterned iconicity in sign language lexicons. *Gesture*, *13*, 287–308.

Peirce, J. W. (2007). PsychoPy–Psychophysics software in Python. *Journal of neuroscience methods*, *162*, 8–13.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

Raviv, L., & Arnon, I. (2018). Systematicity, but not compositionality: Examining the emergence of linguistic structure in children and adults using iterated learning. *Cognition*, *181*, 160–173.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151–164.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. McWhinney, & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 237–263). Hoboken,NJ: Wiley-Blackwell.

Richie, R., Yang, C., & Coppola, M. (2014). Modeling the emergence of lexicons in homesign systems. *Topics in Cognitive Science*, *6*, 183–95.

Sandler, W., Meir, I., Padden, C., & Aronoff, M. (2005). The emergence of grammar: systematic structure in a new language. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 2661–2665.

Schouwstra, M., & de Swart, H. (2014). The semantic origins of word order. *Cognition*, *131*, 431–6.

Senghas, A., & Coppola, M. (2001). Children creating language: how Nicaraguan sign language acquired a spatial grammar. *Psychological Science*, *12*, 323–8.

Senghas, A., Kita, S., & Ozyürek, A. (2004). Children creating core properties of language: evidence from an emerging sign language in Nicaragua. *Science*, *305*, 1779–82.

Silvey, C., Kirby, S., & Smith, K. (2014). Word Meanings Evolve to Selectively Preserve Distinctions on Salient Dimensions. *Cognitive science*, (pp. 1–15).

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, *49*, 370–407.

Smith, K., Abramova, K., Cartmill, E., & Kirby, S. (2017a). The experimental study of sign language emergence, . Manuscript in preparation.

Smith, K., Perfors, A., Feher, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017b). Language learning , language use , and the evolution of linguistic variation. *Philosophical transactions of the Royal Society B*, *372*, 20160051.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*, 444–9.

Sulik, J. (2018). Cognitive mechanisms for inferring the meaning of novel signals during symbolisation. *Plos one*, *13*, e0189540.

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

Theisen-White, C., Kirby, S., & Oberlander, J. (2011). Integrating the horizontal and vertical cultural transmission of novel communication systems. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Expanding the Space of Cognitive Science: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* 2010 (pp. 956–961). Austin, Texas: Cognitive Science Society.

Tkachman, O., & Sandler, W. (2013). The noun–verb distinction in two young sign languages. *Gesture*, *13*, 253–286.

Verhoef, T., Kirby, S., & de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, *43*, 57–68.

Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition*, *176*, 15–30.

# 9. Supplementary materials

## 9.1. Video examples

The gesture videos from all experiments can be found in the University of Edinburgh's DataStore, at http://dx.doi.org/10.7488/ds/2447.

## 9.2. Analyses

Data files and Jupyter notebooks detailing all analyses can be accessed via the Open Science Framework at https://osf.io/psxz6/.

## 9.3. Gesture shape distributions

Figure 19 illustrates the number of gestures against their frequencies for each generation of each chain in experiment one. Unique gesture shapes described by the coding scheme are given on the x-axis, at each generation, and their frequency is shown on the y-axis. These charts demonstrate the similar trajectories that each chain follows in regularising the gesture shapes they use, which is measured in the main text using entropy. The pathways of this process in each chain are remarkably similar. We suggest that the ways in which participants create regular gesture sets is limited; they begin with a larger number of different gestures, and through communication and transmission, settle on particular shapes, eliminating competing gestures. This leads to a smaller pool of gestures being re-used more frequently, and can occur regardless of the particular gestures used.

## 9.4. Communicative accuracy

We noted the accuracy at each trial, i.e. whether the matcher guessed the meaning correctly from their partner's gesture. Figure 20 shows the percentage of correct responses at each generation or round across conditions where interaction is present. Accuracy cannot be measured for the transmission-only condition, as there is no gesture interpretation involved in the testing stage of the experiment.

Logistic mixed effects models analysed the effect of generation and condition on communicative accuracy. Chain (or pair) and target meaning were included as random effects with random intercepts, and a random slope of generation (or round) was included for chain (or pair). The random effects for participant were nested within chains. Transmission + interaction was used as the baseline condition, here and throughout this

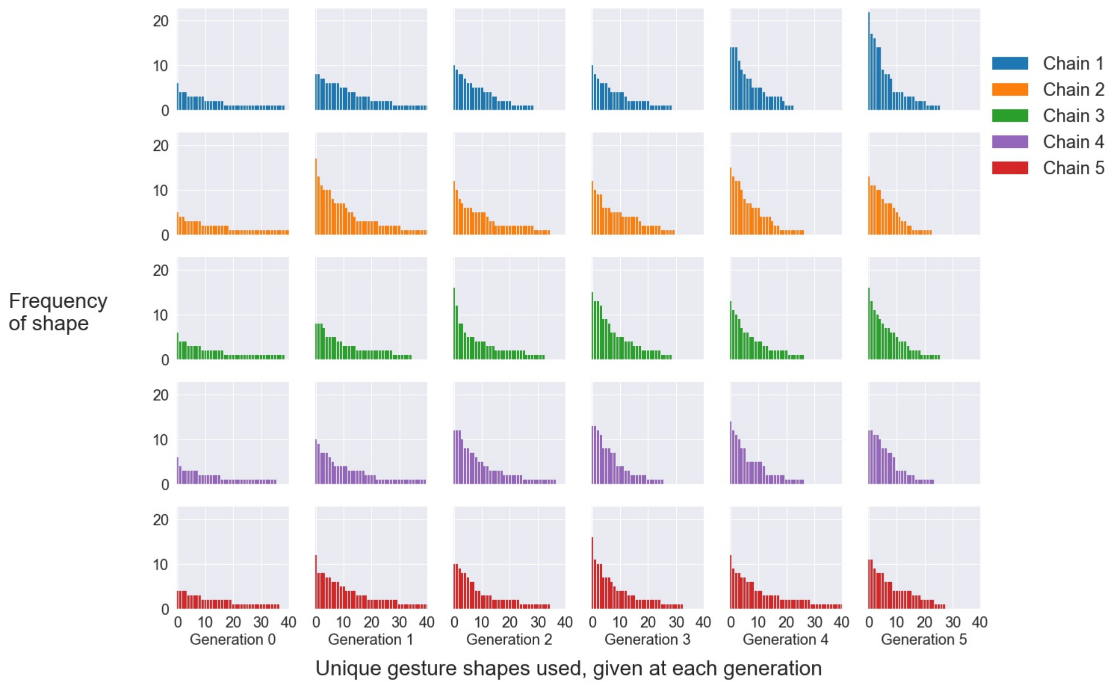
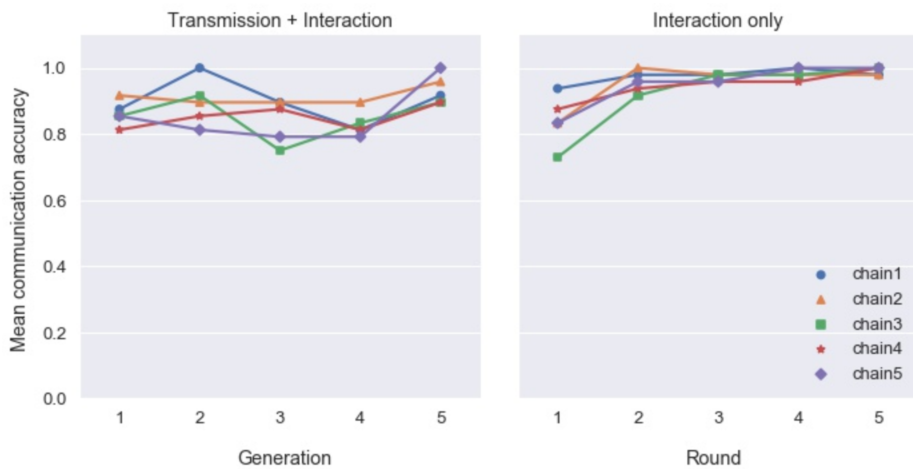


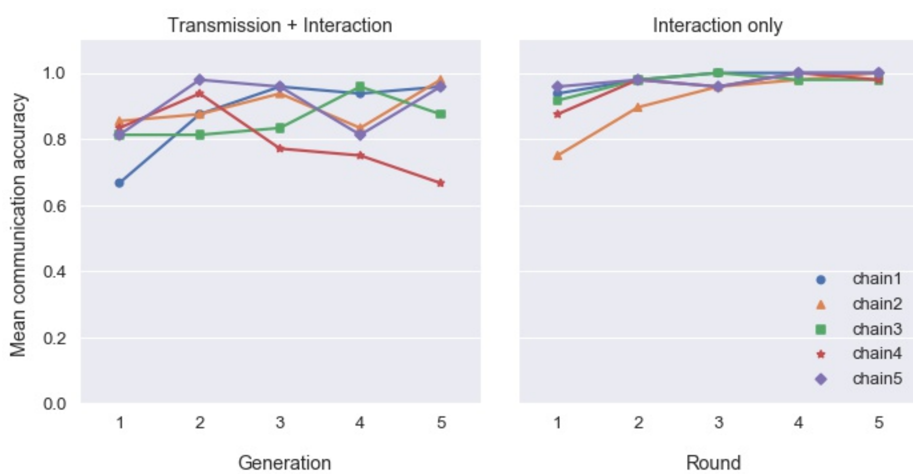

Figure 19: Frequency of unique gesture shapes at each generation, for each chain in experiment 1. Over generations in each chain, participants increasingly re-use the same atomic gestures, from a smaller pool of gestures. For example, at generation 0 seed participants use a wide range of different gestures in low frequencies. By generation 5, participants re-use the same gestures in higher frequencies, and use fewer unique gestures.

section. The random effect structure specified here is used for all subsequent models in this section. One model compared the transmission + interaction data from experiment 1 with the interaction-only data from experiment 2 (both with explicit interactional pressures); another model compared the two corresponding conditions in experiment 3 (without explicit interactional pressures). Results from each model are given in table 6.

We find no main effect of generation for the transmission + interaction condition, though for each comparison we find a significant interaction between generation and condition, suggesting that accuracy in the interaction-only conditions does increase over generations, in each case. These results suggest that gestures in the interaction-only condition fulfil a primary purpose of facilitating communication between partners. The




(a)




(b)

Figure 20: Communicative accuracy for a) experiment 1 (transmission + interaction) and the interaction-only condition of experiment 2, and b) the two corresponding conditions of experiment 3. Accuracy does not increase over generations in the transmission + interaction conditions, but demonstrates an increase over generations in the interaction-only conditions.

| | $\beta$ | $SE$ | $z$ | $p$ |
|---|---|---|---|---|
| **Model 1** | | | | |
| generation | 0.07 | 0.08 | 0.84 | 0.40 |
| condition | 0.03 | 0.26 | 0.10 | 0.92 |
| condition * generation | 0.80 | 0.15 | 5.30 | <0.001*** |
| **Model 2** | | | | |
| generation | 0.19 | 0.14 | 1.35 | 0.18 |
| condition | 0.59 | 0.30 | 1.98 | 0.05 |
| condition * generation | 0.68 | 0.16 | 4.16 | <0.001*** |

Table 6: Model summary for logistic mixed effects models analysing the effect of generation and condition on communicative accuracy. Model 1 compares transmission + interaction data from experiment 1 and interaction-only data from experiment 2. Model 2 compares corresponding conditions in experiment 3. For each fixed effect, we give the beta value, the standard error, the $z$ statistic and the p-value.

lack of change over generations for the transmission + interaction conditions may illustrate the trade off between transmission + interaction; in this case, a drive for communicative accuracy trades off with a pressure for learnable gestures. Note, however, that communicative accuracy is still high in the transmission + interaction conditions (mean accuracy: experiment 1 = 87.3%, experiment 3 = 86.5%).

### *9.5. Within- and between-generation similarity*

To assess alignment and learnability, we measured gesture similarity in two ways, using the same measure. We measured the similarity of a participant's gestures to (a) the gestures of their partner in communication, and, (b) the the gestures of their training model. We will call the former within-generation similarity, and the latter between-generation similarity. Within-generation similarity offers a measure of alignment between communication partners, and between-generation similarity gives an indication of how learnable gestures are, by measuring how well participants reproduce the gestures they see in training.

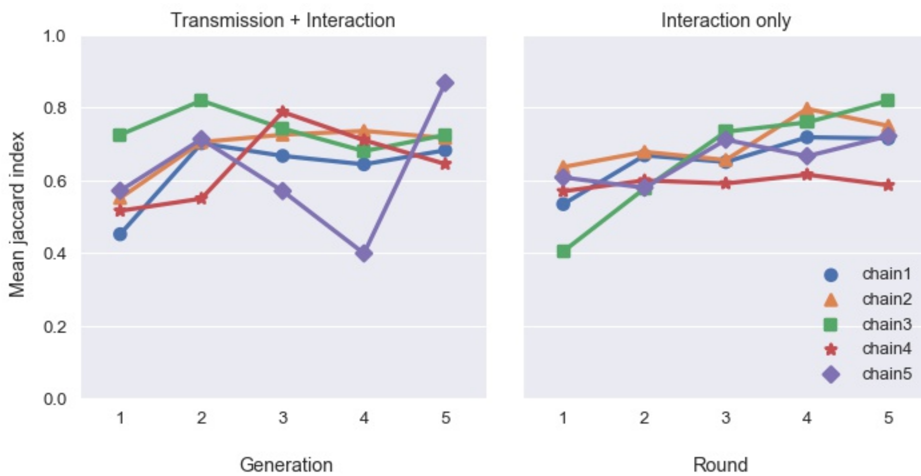
Gesture similarity is based on gesture coding strings and is given as the Jaccard index, a similarity measure defined as,

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

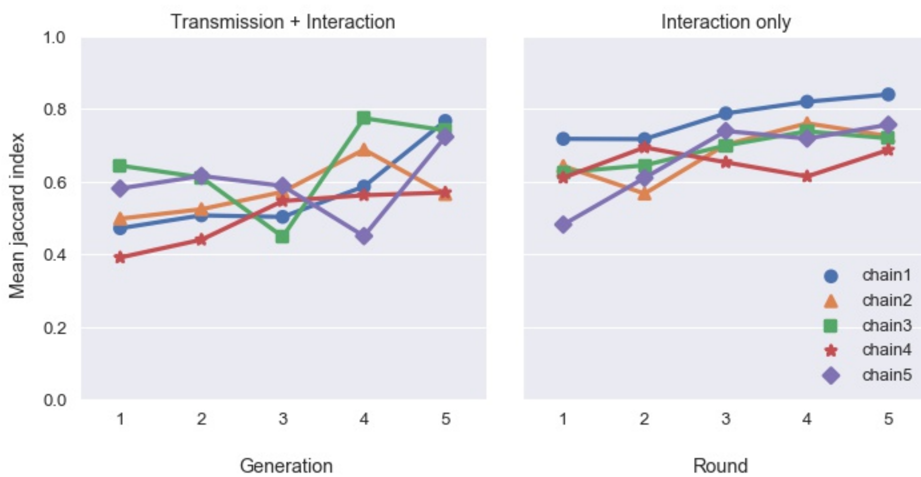
denoting the intersection of two sets (only what is shared between them) divided by the union of the two sets (all unique elements across both sets). Within-generation similarity was calculated for experiment 1 (transmission + interaction) and the interaction-only condition in experiment 2, as well as for both corresponding conditions of experiment 3. Within-generation similarity is not measured in the transmission-only condition, where there is only one participant per generation. Similarly, between-generation similarity was calculated for experiment 1 (transmission + interaction), the transmission-only condition in experiment 2, and the transmission + interaction condition of experiment 3. We did not measure between-generation similarity for the interaction-only conditions in experiments 2 and 3, as no new learners are introduced, and no further training takes place.

*Within-generation similarity.* Linear mixed effects models analysed the effect of generation and condition on alignment scores. As with the models measuring accuracy, one model compared alignment between the transmission + interaction data from experiment 1 and the interaction-only data from experiment 2 and another model compared alignment in the conditions of experiment 3. Model results are shown in table 7; figure 21 illustrates results from both comparisons. With both comparisons, we find an increase in alignment over generations, and no interaction between condition and generation. With or without explicit interactional pressures, communicating participants become more aligned over generations or rounds, suggesting that gestures become increasingly conventionalised when used in interaction.

*Between-generation similarity.* The model structure described above was used to analyse the effect of generation and condition on between-generation similarity. One model compared learnability in the transmission + interaction data from experiment 1 with the transmission-only data from experiment 2; another model compared learnability in the transmission + interaction data from experiment 3 with the transmission-only data from experiment 2. Model results are shown in table 8.

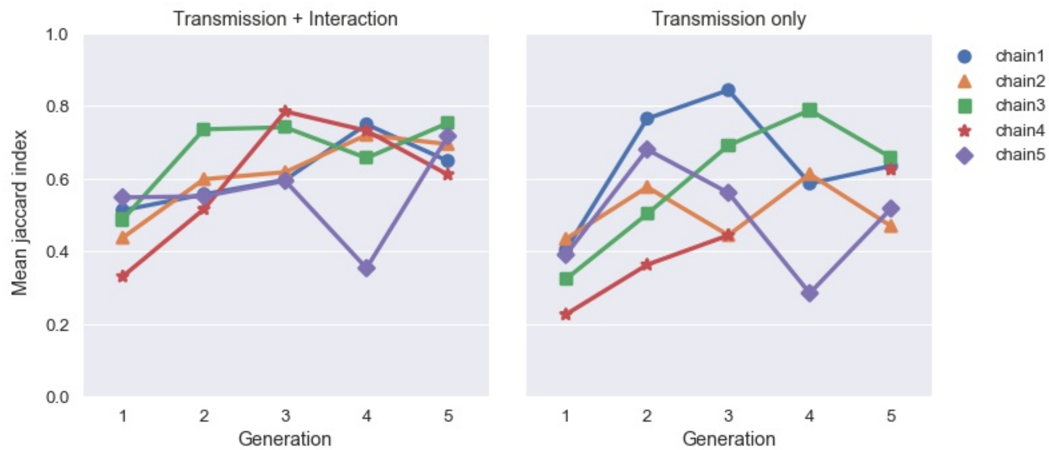



(a)

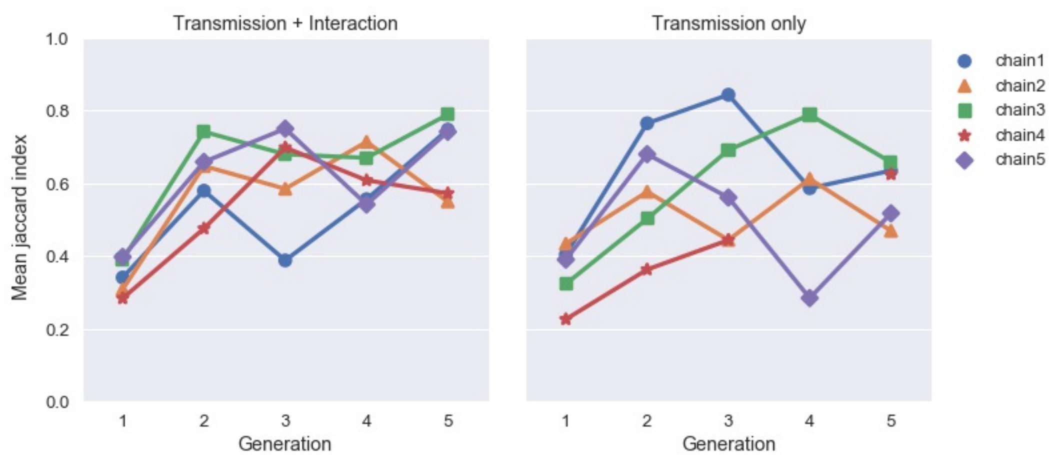



(b)

Figure 21: Within-generation similarity for a) experiment 1 (transmission + interaction) and the interaction-only condition of experiment 2, and b) the two corresponding conditions of experiment 3. Alignment increases over generations, with no significant difference between conditions in both cases.




(a)




(b)

Figure 22: Between-generation similarity for a) experiment 1 (transmission + interaction) and the transmission-only condition of experiment 2, and b) the transmission + interaction condition from experiment 3, and the transmission-only condition from experiment 2. Learnability increases over generations, with no significant difference between conditions in both cases.

| | $\beta$ | $SE$ | $t$ | $p$ |
|---|---|---|---|---|
| **Model 1** | | | | |
| generation | 0.03 | 0.01 | 2.63 | 0.01* |
| condition | -0.04 | 0.03 | -1.22 | 0.22 |
| condition * generation | 0.02 | 0.01 | 1.13 | 0.25 |
| **Model 2** | | | | |
| generation | 0.04 | 0.009 | 4.20 | <0.001*** |
| condition | -0.12 | 0.03 | -4.10 | <0.001*** |
| condition * generation | -0.004 | 0.01 | -0.33 | 0.74 |

Table 7: Model summary for within-generation similarity. Model 1 compares transmission + interaction data from experiment 1 and interaction-only data from experiment 2. Model 2 compares corresponding conditions in experiment 3. For each fixed effect, we give the beta value, the standard error, the *t* statistic and the p-value.

Both comparisons demonstrated an increase in between-generation similarity over generations, and found no interaction between condition and generation. Thus, where transmission is present, we find an increase in learnability over generations, with participants better able to reproduce the gestures they have learnt in training.

| | $\beta$ | $SE$ | $t$ | $p$ |
|---|---|---|---|---|
| **Model 1** | | | | |
| generation | 0.05 | 0.02 | 3.27 | 0.02* |
| condition | -0.06 | 0.05 | -1.28 | 0.21 |
| condition * generation | -0.005 | 0.02 | -0.24 | 0.81 |
| **Model 2** | | | | |
| generation | 0.07 | 0.01 | 4.92 | $<$0.001*** |
| condition | 0.003 | 0.06 | 0.06 | 0.96 |
| condition * generation | -0.02 | 0.02 | -0.99 | 0.33 |

Table 8: Model summary for between-generation similarity. Model 1 compares transmission + interaction data from experiment 1 and transmission-only data from experiment 2. Model 2 compares corresponding conditions in experiment 3. For each fixed effect, we give the beta value, the standard error, the *t* statistic and the p-value.