Multiscale kinematic analysis reveals structural properties of change in evolving manual languages in the lab

Wim Pouw[1], Mark Dingemanse[1,3], Yasamin Motamedi[4], & Asli Ozyurek[1,2,3]

[1] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen

[2] Institute for Psycholinguistics, Max Planck Nijmegen

[3] Center for Language Studies, Radboud University Nijmegen

[4] Language and Cognition Lab, University College London

Author Note: Correspondence concerning this article should be addressed to Wim Pouw, Montessorilaan 3, 6525 HR Nijmegen. E-mail: w.pouw@psych.ru.nl. This work is supported by a Donders Fellowship awarded to Wim Pouw and Asli Ozyurek and is supported by the Language in Interaction consortium project 'Communicative Alignment in Brain & Behavior' (CABB).

Open data: All data and analysis supporting this paper can be found on https://github.com/WimPouw/GestureNetworksEvolving. This manuscript has been written in Rmarkdown, a code-embedded version (.Rmd) can be found on our Github page.

Abstract

Reverse engineering how language emerged is a daunting interdisciplinary project. Experimental cognitive science has contributed to this effort by eliciting in the lab constraints likely playing a role for language emergence; constraints such as iterated transmission of communicative tokens between agents. Since such constraints played out over long phylogenetic time and involved vast populations, a crucial challenge for iterated language learning paradigms is to extend its limits. In the current approach we perform a multiscale quantification of kinematic changes of an evolving silent gesture system. Silent gestures consist of complex multi-articulatory movement that have so far proven elusive to quantify in a structural and reproducable way, and is primarily studied through human coders meticulously interpreting the referential content of gestures. Here we reanalyzed video data from a silent gesture iterated learning experiment (Motamedi et al. 2019), which originally showed increases in systematicity of gestural form over language transmissions. We applied a signal-based approach, first utilizing computer vision techniques to quantify kinematics from videodata. Then we performed a multiscale kinematic analysis showing that over generations of language users, silent gestures became more efficient and less complex in their kinematics. We further detect systematicity of the communicative tokens's interrelations which proved itself as a proxy of systematicity obtained via human observation data. Thereby we demonstrate the potential for a signal-based approach of language evolution in complex multi-articulatory communication.

Keywords: language evolution, silent gesture, kinematics, systematicity

Word count (in text): 9412

Introduction

There is an ongoing scientific effort to unveil the historical and/or necessary constraints that allow(ed) for the emergence of human language (e.g., Bickerton, 2009; Deutscher, 2005; McNeilage, 2008; Tomasello, 2008). An important approach within this enterprise comes from experimental cognitive science (Scott-phillips & Kirby, 2010). In this approach interactive communication processes likely to have contributed to language emergence are simulated in the lab with human (e.g., Kirby, Cornish, & Smith, 2008) and sometimes non-human primate subjects (e.g., Claidière, Smith, Kirby, & Fagot, 2014), who are tasked to employ a variety of communicative systems (Cornish, Dale, Kirby, & Christiansen, 2017; Ravignani, Delgado, & Kirby, 2016; Verhoef, Kirby, & de Boer, 2016). Some communicative systems are more transparent than others however, and a standing challenge for this field is to systematically quantify continuous multidimensional communicative signals *as communicative systems*, but without reducing (or enriching) such signals to discrete meanings by top-down judgments of human coders. Here we show that complex events such as manual- and head-movement gestures can be studied from the buttom-up *as continuous events* by probing the interrelationships with other gestures, in relation to the changing movement properties of the gestures themselves.

The experimental cognitive science approach to language evolution often involves agents learning a novel set of signals which is iteratively transmitted to later generations (iterated learning) or also used in communication by later generations (iterated learning + communication). Over many cycles of learning and use, the signals are affected by various transmission biases (e.g., Christiansen & Chater, 2016; Enfield, 2016). Processes of iterated learning and communication can simulate how structural properties such as systematicitity, learnability, and compositionality evolve from more simpler communication systems — a process that must have occurred in human language evolution too (Bickerton, 2009). In such simulations items undergoing cultural evolution abide by population dynamic constraints such as historicity (the system is constrained by past contingencies) and adaptivity (the system is able to tweak itself in service of its informative goals). Such population dynamics must have played out over long temporal and vast population scales, but through these iterated learning paradigms such processes are to some limited degree brought under experimental control. A current challenge is to

extend the limits of such paradigms and study how the same constraints can give rise to novel emergent structure at larger scales of interaction (e.g., Lou-Magnuson & Onnis, 2018; Lupyan & Dale, 2010; Raviv, Meyer, & Lev-Ari, 2019). Such extension requires currently absent automated methods for the study of complex dynamic signals as part of communicative systems.

The current report showcases a signal-based approach for the study of kinematic communication systems — in this case silent gestures, i.e., manual communicative movements produced in the absence of speech. As we review below, silent gestures are a promising locus for studying the cultural evolution of signs and signalling. But they are also challenging to study given their continuous and complex (multi-)articulatory nature. Here we build on data from a recent iterated learning paradigm with silent gestures, wherein users reproduced communicative gestures within chains of 5 iterated generations (Motamedi, Schouwstra, Smith, Culbertson, & Kirby, 2019). With computer vision (Cao et al., 2017) we obtained motion traces of manual- and head gestures. Subsequently we performed 'gesture network analysis' (Pouw & Dixon, 2019), which is a procedure that combines bivariate time series analysis (Dynamic Time Warping) with network analysis and visualization. Next to reporting kinematic changes indicative of communicative efficiency, we show through gesture network analysis that there is an emergence of systematicity, which approximates systematicity obtained from the human-coded gesture content. We further show how such systematicity is reflected in the reduction of kinematic complexity of gesture utterances as the communicative system evolves. As such this multi-scale analysis is able to relate form level characteristics of gesture utterances with higher level characterizations of systematicity, breaking ground for a quantitative study of manual- and whole body movements as communicative systems (see e.g., Sandler, 2018).

Language evolution and silent gesture

Some scholars of language evolution hold that human language must have started in the manual or whole-body modality (Corballis, 2002; Donald, 1991; Tomasello, 2008) while others have suggested that the manual modality and vocal systems have co-evolved (e.g., Levinson & Holler, 2014; Kendon, 2017). Such opposing views are nevertheless united by their conviction that human language is firmly rooted in manual

communication, as evidenced by the pervasiveness of co-speech gesture and the ease of for humans to instantiate language in the manual modality.

Given the relative scarcity of people who master a sign language, silent gesture is especially interesting tool for studying language evolution de novo. It has for example been shown that syntactic conventions in a spoken language are not necessarily reproduced cross-modally in silent gestures, rather hearing participants' silent gestures will follow novel syntactic conventions independent of spoken language (Goldin-Meadow, So, Özyürek, & Mylander, 2008; Schouwstra, 2017). Thus, silent gestures are to some degree *authentically* produced and it allows researchers to tap into biases that shape communication while reducing the influence of existing linguistic knowledge.

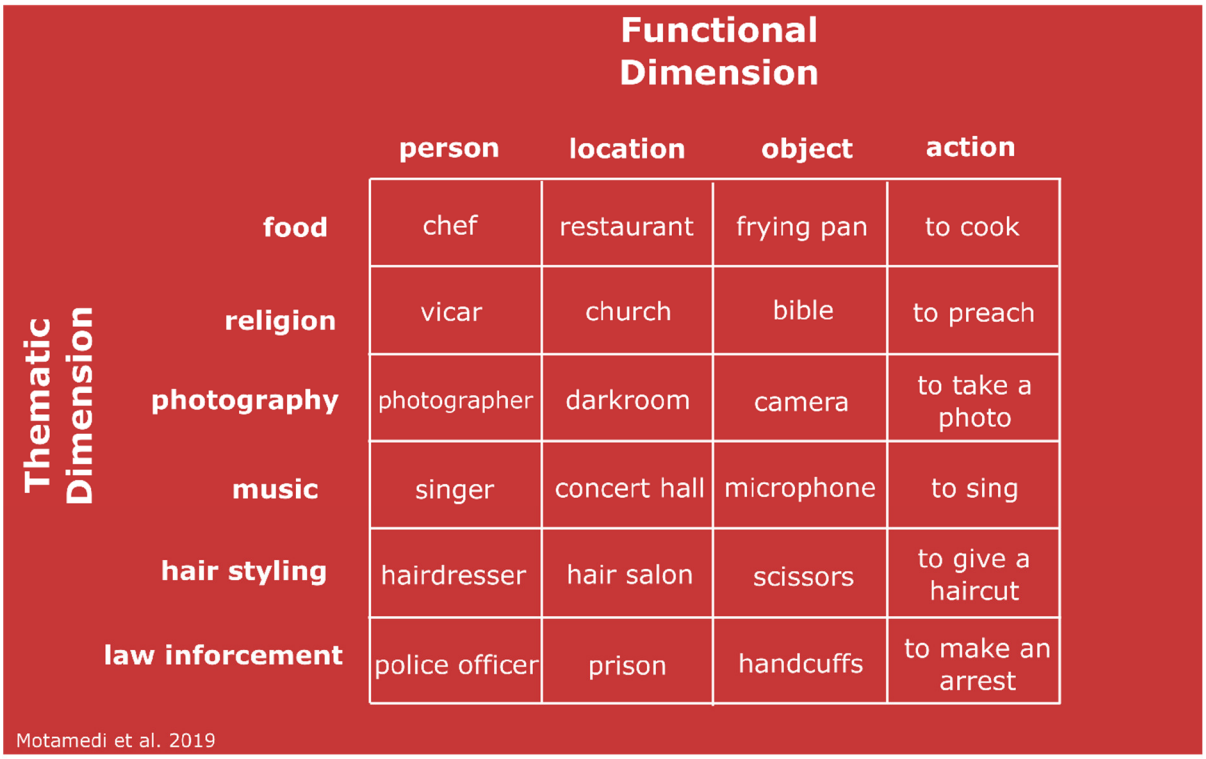
Gestures naturally afford visual-motor mappings to referents, that is they tend towards iconic presentation (e.g., Ortega, Schiefner, & Ozyurek, 2019; Ortega & Özyürek, 2020a). The manual modality is of course not unique in this, as spoken languages show plenty of iconicity (Dingemanse et al., 2015), but hand movements are unique in the flexible way they can present visual iconic mappings and the degree to which they do so. It has been reported that in some sign languages communicative load can be carried to much greater extent by iconic means, which would otherwise need to be carried by other linguistic innovations such as combinatorial phonology (Aronoff, Meir, Padden, & Sandler, 2008; Slonimska, Özyürek, & Capirci, 2020). Indeed, gesture-first theories emphasize that there is a natural grounding of gestures in routine behaviors such as manual action with the environment, allowing perceivers to more easily recognize these as communicatively relevant movements, and enabling the development of communicative conventions.

While gestures have their natural tendencies of expression, they have been found to flexibly and quickly adapt to the social context. For example, in dyadic social interaction, repeated gestural referrals to an object or a picture will lead to those gestures becoming more reduced in size (Gerwing & Bavelas, 2004; Namboodiripad, Lenzen, Lepic, & Verhoef, 2016). This is comparable to research in 'pictionary' paradigms where a concept is drawn out and to be interpreted by another player. After repeated trials of drawing, a reduction of the drawings' complexity is observed, with smaller-sized and less iconic drawings as a result, while communicative accuracy increases over time (Fay, Garrod, Roberts, & Swoboda, 2010; Garrod, Fay, Lee,

Oberlander, & Macleod, 2007). Though, drifts from less or more iconicity are not fixed processes. When interaction between people is opened up, a whole new suit of social affordances arise. For example, while gestures may reduce in size and iconicity when some common ground is established, at any moment an interlocutor may request a clarification, soliciting large and iconic gestures per implicitly requested (Bavelas, Gerwing, Sutton, & Prevost, 2008; Holler & Wilkin, 2011). In such moments of interactional repair, common ground is calibrated and re-established. These and many other interactive and dialectical affordances turn out to be of central importance for smooth everyday language use (Dingemanse, Roberts, et al., 2015), and according to cultural evolutionary accounts of language, such local-scale processes of interaction and transmission between communicators are crucial for the emergence of any linguistic system (e.g., Enfield, 2016; Kirby & Christiansen, 2003; Kirby, Griffiths, & Smith, 2014; Raviv et al., 2019). A key question that drives cultural evolution research is which particular interactive constraints produce pressures for a certain communication system to adapt in one way or another, and how effective solutions are negotiated at the possible expense of other communicatively efficient solutions (Dingemanse et al., 2015).

In a recent iterated learning study with silent gesture (Motamedi et al., 2019) two such possible constraints, transmission across generations and communication within generations, were studied simultaneously as well as separately. Learning occurred with a set of silent gesture-concept mappings (i.e., communicative tokens) communicated through five iterations of vertical transmissions, where gestures were transmitted from one participant to-be-reproduced by the next participant. Or, tokens would be communicated through five horizontal interactions in a director-matcher type task. These constraints - interaction and transmission - were first studied in combination in experiment 1, which is the focus for the current paper. An important aspect of the study was that every concept was characterisable along two dimensions: theme (e.g., food, religion) and function (e.g., person, location) (see Figure 1).

Figure 1. Concepts to be conveyed in gesture in Motamedi et al. 2019



These dimensions provided possible axes for compressibility of the communicative tokens. After all, by combining 10 unique gestures one can pick out any referent (e.g., "to make an arrest") from the 24 token meaning space, one gesture marking the functional category (e.g., "action") and another gesture for the theme category ("justice"). To exemplify further, once confronted with communicating 24 meanings one can invent 24 unique gesture utterances, such as in the following videos for "to sing" (https://osf.io/d8srx/) and "singer" (https://osf.io/974ke/), which is challenging to do since they are both very much related. However, one can also start differentiating by functional category such that "microphone" is preceded by a general object marking gesture ("https://osf.io/r3gcp/") and "singer" is preceded by a general person marking gesture ("https://osf.io/ex4tv/"), and then followed by the same thematic marking gesture conveying "music". Not only do these general functional markers aid the disambiguation of related meanings, this expressive invention also allows for a systematic reemployment of functional markings for the whole meaning space through compositionality. Once you invent 4 functional marker gestures, and 6 thematic marker gestures, you can systematically recombine these to convey 24 meanings. The communicative system then has compressed its information density from 24 information

units to 10 information units. Motamedi and colleagues (2019) indeed observed such signs of compression of the meaning space as the system developed. In early iterations of learning, large-sized iconic enactments were the most common way of gesturally depicting the referents. However, the occurrences of functional markers increased over generations, which represented meaning components reused across gestures. This kind of functional marking mainly targeted the thematic and function categories.

With meticulous hand coding of the different referential components of each silent gesture, it could further be quantitatively tested whether there was indeed systematicity emerging. The gesture coding included information about form of a particular gesture segment, such as the number of manual articulators used (1 or 2 hands), as well as the referential target of the gesture (e.g., hat; pan; turn page). Based on the full sequences of the referential components that were uniquely expressed in each gesture, entropy was computed, which expresses compressibility of the gesture content, i.e., the amount of information that is needed to compress the signal. When a lot of referential components in the gesture utterance recurs between other gestures, the system has a more simple structure and indicates systematic reuse of gestural components (e.g., Gibson et al., 2019). Dovetailing with the qualitative observations and other studies in this field (e.g., Verhoef et al., 2016), it was found that gesture-component entropy decreased over the generations. Furthermore, the gestures were coded for the amount of marking for the functional category, and this showed that such gestures occurred more often at later generations. Finally, average gesture duration - as a measure of communicative efficiency - did not reliably change over the generations, which ran counter to predictions that more mature communication systems tend towards maximal efficiency (Gibson et al., 2019).

These results obtained in the lab resonate with findings from homesign (e.g., Haviland, 2013) and emerging sign languages (Senghas, Kita, & Özyürek, 2004). For example, it has been shown that in the expression of motion events first generation signers of Nicaraguan sign language performed more holistic presentations of path and manner, while in following generations manner and path were segmented. Such segmentations affords novel combinatoriality and therefore increases generativity of a language. It expresses the meaning space with fewer means similar to how participants

studied by Motamedi et al. (2019) started to compress the meaning space by developing ways mark functional status across referents (e.g., “agent”, “action”).

Current approach: Going beyond the state of the art

So far research on linguistic properties of manual or whole-body gesture has been based on human coding and focused on semantic rather than form analysis. Often this is theoretically well justified because the kinematic signal — similar to acoustics in speech — does not exhaust the content of the signal. That is, although gesture can be objectively rendered by its kinematics — i.e., rendered as a bodily posture in movement through space — a gesture’s meaning is not contained in the kinematics as such. A communicative context and a community of language users is needed to decide on such meanings, with the human coder acting as the representative. However, there are many alien communication systems (e.g., bidsong) from which we can detect languagelike properties. Indeed, the form-level systematicities, such as in acoustics, or for that matter kinematics, can be revealing of linguistic structure, and the emergence of such structure has indeed been found in many different kinds of communicative signals, such as whistling signals controlled by a slider (Verhoef et al., 2016), drumming sequences (Ravignani et al., 2016), letter sequences (Cornish et al., 2017), and a wide range of animal vocalizations (Engesser & Townsend, 2019).

A pressing challenge for applying a similar approach to silent gesture is how to quantify systemic changes from continuous and complex multi-articulatory body movements. While there has been progress in quantifying form similarity between silent gestures (e.g., Namboodiripad et al., 2016; Sato, Schouwstra, Flaherty, & Kirby, 2020), a standing challenge is how to understand such kinematic events at higher levels of description, which involves the study of communicative tokens in the context of the larger system they may be part of. Namely, it is one thing to show that a gesture changes kinematically over multiple generations of producers, it is another thing to show that kinematic changes of such a gesture are systematically related to the kinematics of other gestures produced. Only the latter analysis can reveal that gestures’ form evolves as a communicative system.

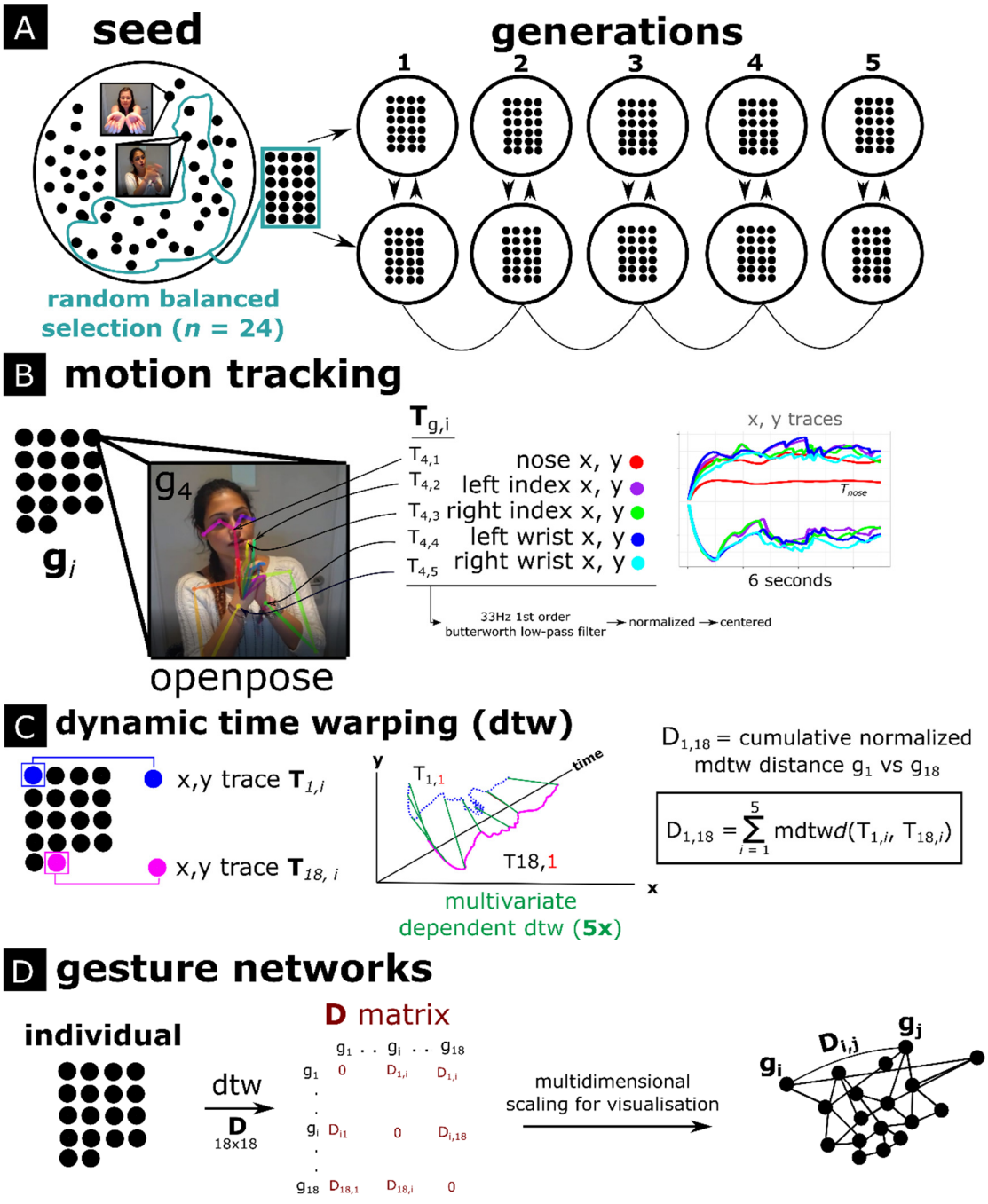
Here we address this challenge of relating dynamic multi-segmented kinematics with the possible systematicity emerging at the level of meaning between gesture events.

To this end, we first applied computer vision techniques (Cao et al., 2017) to extract human movement traces from video data, and submitted these multidimensional time series to gesture network analysis (Pouw & Dixon, 2019). This approach uses a well-known similarity comparison of time series (e.g., Verhoef et al., 2016; Sato et al., 2020), which is then leveraged to produce all possible comparisons between communicative tokens. Having mapped all interrelationships of communicative tokens, which is essentially a network, you can then characterize the topology of such networks to understand systematic relations within the communicative system. Specifically, we assessed entropy of the gesture network connections (containing temporal, spatial, and segmental interrelations between gestures), which much like the original study assesses the compressibility of the interrelationships between communicative tokens. We therefore predicted that network entropy based on continuous kinematic data, would approximate entropy based on discrete human gesture codings. Once you have evidence for systematic changes in the communicative system (higher scale) the next step will be a kinematic analysis of the communicative tokens (lower scale). Therefore, we further report how temporal, spatial, and segmental properties of gesture kinematics changed as the communicative system evolved, which we relate to changes at the system level. The study of Motamedi et al. (2019) provides an ideal ground-truth for the current signal-based approach, as gesture form and its information units has been extensively documented in a transparent way. As such, the current data provides a platform to launch a new approach for going beyond discrete detection of gesture by coders at the level of meaning, to a continuous analysis of language in movement at multiple scales of analysis, yielding new insights on how gesture kinematics changes *as communicative systems* during iterated learning. The current research is the litmus test of this multiscale approach.

Method

We will follow a bottom-up approach to the study of kinematics as communicative systems, by first assessing possible systematic interrelationships in kinematic patterns of gestures through gesture network analysis (step 1).Gesture network analysis aims to target structural properties existing on the system level, studying the relations between tokens rather than the form or content of those tokens. However, it is equally important to understand what specific changes occur in the kinematics of the gestures, as such changes might predict changes on the system level. Therefore, the next step will be a fine-grained kinematic analysis (step 2) where we will overview the kinematic analysis of the gestures themselves based on what we think are relevant dimensions for articulatory complexity that might have given rise to results obtained in step 1. For step 1, Figure 2 shows the general overview of the gesture network analysis procedure for this experiment. We will discuss each step in this procedure in the following sections, and discuss our main network measures. We invest extra space for providing quantitative checks to motivate our particular measurement choices against possible alternative choices.

Figure 2. General method gesture network analysis



Note Figure 2. The general procedure is shown for the current gesture network analysis. A) shows the original experiment setup (Motamedi et al., 2019), where a seed set of 24 gestures was randomly selected for each chain containing five generations. Seed gestures were used to train the first generation of each chain; subsequently, gestures from the previous generation were used as training data. Participants then communicated gesturally about the same concepts. B) For our analysis we first performed video-based motion tracking with OpenPose (Cao et al., 2017) to extract relevant 2D movement traces (T_i) of the

nose, the wrists and index fingers. C) For each gesture comparison within a gesture set, the time series were then submitted in to a Dynamic Time Warping procedure where we computed for each body part a multivariate normalized distance measure, repeated for all body parts and summed, resulting in one overall distance measure D for each gesture comparison. D) All distance measures were saved into a matrix **D** containing all gesture comparisons $D_{i,j}$ within the comparison set, resulting in a 24x24 distance matrix. The distance matrix can be visualized as a fully connected weighted graph through multidimensional scaling, such that nodes indicate gesture utterances and the distance (or weight) between gesture nodes representing the 'D' measure, indicating dissimilarity.

Participant, design, & procedure of the original study (experiment 1). Here we discuss the setup of the experiment which generated the data we reanalyzed (for more detailed information see Motamedi et al., 2019).

A seed gesture set was created with 48 pre-study participants who each depicted 1 out of 24 concepts. Thus for each concept there were two seed gestures performed by unique pre-study participants. Given that pre-study participants only produced one gesture, they were isolated from the other concepts that comprised the meaning space.

For the main experiment (exp. 1) 50 right-handed English-speaking non-signing participants were recruited. They were allocated pairwise to one of 5 iteration chains. Participants were first shown a balanced subset of 24 unique seed gestures. These chain-specific seed gesture sets will be referred to as generation 0, which were followed by generations 1 through 5. In the training phase, gestures were presented in random order and participants were asked to identify the meaning of the gesture from the 24-item meaning spaces, followed by feedback about their performance. They were then asked to self-record their own copy the gesture. Participants trained with a subset of 18 items (out of 24), and completed two rounds of training.

In the testing phase, participants took turns as director and matcher to gesturally communicate (withou using speech) and interpret items in the meaning space, with feedback following each trial. This director-matcher routine was repeated until both participants communicated all 24 meanings. Subsequent generations were initiated with new dyads whose training set was the gestures from one randomly selected participant from the prior generation.

The recorded videos of the seed gestures and the gesture utterances participants produced in the testing phases are the data we use here. This means that we have 50

participants conveying 24 concepts = 1200 gesture videos belonging to generations 1-5, and 48 seed gesture videos with each concept conveyed by 2 seed participants (i.e., 48 seed participants).

Motion tracking. Motion tracking was performed on each video recording with a sampling rate of 30Hz. To extract movement traces, we used OpenPose (Cao et al., 2017), which is a pre-trained deep neural network approach for estimating human poses from video data (for a tutorial see Pouw & Trujillo, 2019). We selected keypoints that were most likely to cover the gross variability in gestural utterances: positional x (horizontal) and y (vertical) movement traces belonging to left- and right index fingers, wrists, as well as the nose. For all position traces and its derivatives, we applied 1st order 30Hz low-pass Butterworth filter to smooth out high-frequency jitters having to do with sampling noise. We z-normalized and mean-centered position traces for each video to ensure that differences between subjects (e.g., body size) and within-subject differences in camera position at the start of the recording were inconsequential for our measurements.

Dynamic Time Warping (DTW). DTW is a common signal processing algorithm to quantify similarity between temporally ordered signals (Giorgino, 2009; Mueen & Keogh, 2016; Muller, 2007). The algorithm performs a matching procedure between two time series by maximally realigning (warping) nearest values in time while preserving order, and comparing their relative distances after this non-linear alignment procedure. The degree that the two time series need to be stretched and warped indicates how dissimilar they are. This dissimilarity is expressed with the DTW distance measure, with a higher distance score for more dissimilar time series and a lower score for more similar time series.

The time series in the current instance are multivariate, as we have a horizontal (x) and vertical (y) positional time-series data. However, DTW is easily generalizable to multivariate data, and can compute its distances in a multidimensional space if required, yielding a multivariate dependent variant of DTW. We opt for a dependent DTW procedure here as x an y positional data are part of a single position coordinate in space. Additionally, we have 6 of these 2-dimensional time series for each body keypoint. To compute a single distance measure between gestures, we computed for each gesture comparison a multivariate dependent DTW Distance measure per keypoint, which was

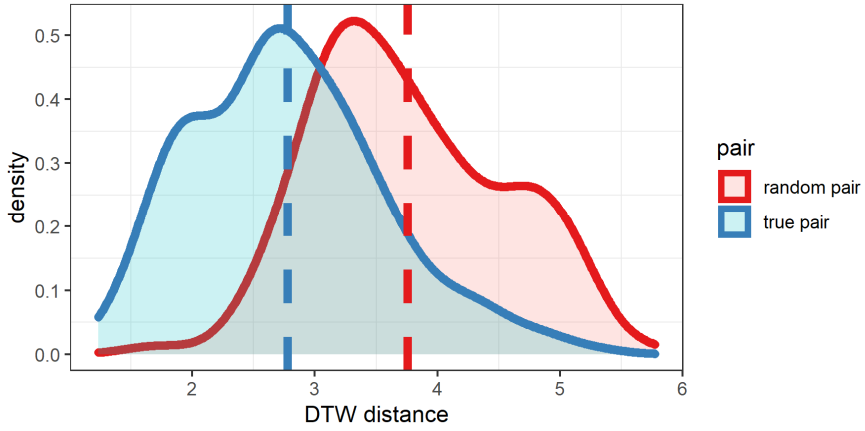
then summed for all keypoint comparisons to obtain a single Distance measure D (illustrated in Figure 2C). The D measure thus reflects a general dissimilarity (higher D) or similarity (lower D) of the whole manual+head movement utterance versus another utterance.

We used the R-package 'DTW' (Giorgino, 2009) to produce the multivariate distances per keypoint. The DTW distance measure was normalized for both time series' length, such that average distances are expressed per unit time, rather than summing distances over time which would yield higher (and biased) distance estimates for longer time series (i.e., longer gesture videos). For further conceptual overview and methodological considerations of our DTW procedure see Pouw and Dixon (2019).

As a demonstration that our D measure reflects actual differences in kinematics, we computed for each individual in each chain the difference between a gesture seed and the gesture that the individual produced to copy it, for generation 1. These "true pairs" must be maximally similar (lower D) as the individual produced their copied gesture short after first exposure in the training phase, which should lead to high faithfulness in reproduction. We contrast this with a false or random comparison of the same gesture in generation 1 with a gesture seed that was neither in the same functional nor thematic category. These false random pairs must be more dissimilar, and should produce higher DTW distances. Figure 3 shows the distributions of the distances observed. DTW distance distributions were reliably different, $t\ (457.97) = 13.45$, $p < .001$, Cohen's $d = 1.25$, for the true pair, $M = 2.78(SD = 0.78)$, as compared to the random pair, $M = 3.75(SD = 0.78)$.

Importantly, we also find that adding head movement trajectory to our D calculation significantly increases false-real pair discriminability as compared when we compute our D measure on only manual keypoints (left/right wrist and index fingers), change in Cohen's $d = 0.37$, change D real vs. false $= 0.33$, $p < .001$. Therefore we conclude that in the current experiment the gesture utterances are also crucially defined by head movements as well. This is an interesting finding in and of itself, and demonstrates the multi-articulatory nature of silent gestures.

Figure 3. Density distributions of D for true pairs and random pairs



Note Figure 3. Density distributions of D are shown for the random versus real pairs. With D based on head-, wrist- and finger movement there is good discriminability between real versus falsely paired gestures, confirming that our approach is tracking gesture similarity well.

Gesture networks. We constructed for each participant (nested in generation and chain), as well as each seed gesture set (seed set belonging to that chain), a distance matrix **D**, containing the continuous D comparisons for each gesture $D_{i,j}$ produced by that participant with each other gesture produced by that participant, yielding a 24x24 distance matrix **D**. The diagonal contains zeros for gesture comparisons that are identical ($D_{i,j} = 0 | i = j$). These characteristics make **D** a weighted symmetric distance matrix.

For each distance matrix we can construct a visual representation of its topology by projecting the distance of gesture tokens on a 2d plane using multidimensional scaling. These networks are fully connected graphs with distances between gesture nodes reflecting our D measure. Such 2d representations are imperfect approximations of the underlying multidimensional data and are only used as visual aids. The uncompressed distance matrices are used to calculate the topological properties, i.e., interrelationships of communicative tokens. We refer to these matrix properties as 'network properties' as these measures are intuitively understood in network terms. For multidimensional scaling, network visualization, and calculations of network entropy we use the R-package 'igraph' (Csárdi, 2019).

Gesture Network Properties

Network entropy. The network entropy measure is almost identical to a classic Shannon entropy calculation, where $Entropy\ H(X) = -\sum p(X)\log p(X)$. The only difference is that our measure is computed on the weights of the networks' edges for each node relative to the shortest path to the other nodes (ie., connections), and then normalized by the number of connections.

Entropy is a measure that quantifies the compressibility of data structures, and has been used to gauge the combinatorial structure of communicative tokens in the field of language evolution (e.g., Verhoef et al., 2016; for theoretical grounding see Gibson et al., 2019). In the original experiment, Motamedi and colleagues (2019) computed entropy from the gesture content codings, which captured recurrent information units between gestures. In our case, entropy quantifies the degree to which there are similar or more diverse edge lengths (i.e., similar/diverse levels of dissimilarity 'D'). If they are more similar, this means lower entropy reflecting that communicative tokens relate in more structural ways to each other. Thus it is important to emphasize here that network entropy gauges in our case how the kinematics interrelationships are compressible (show systematic recurrence), and this is conceptually similar as gauging the systematic recurrence of information units between the human judged gesture content. If this is correct, network entropy of kinematic patterns should scale to entropy based on discrete gesture codings.

To explain entropy with some simple examples: if we have a network where the chance of having an edge length of D = x is 1, then the network connections are fully compressible and we yield an entropy of 0 ($Entropy = -1 * 1 * log(1) = 0$). If there are different edge lengths (increasing the complexity of our network) such that we have a 0.5 chance that $D = x$ and 0.5 chance that $D = y$, then entropy goes up, $Entropy = (-1 * 0.5 * log(0.5)) + (-1 * 0.5 * log(0.5)) = 0.68$ (remember that the log of a fraction becomes a negative number, that is why the result is multiplied by -1 at the start of the formula). Note further that when the system is so diverse that there is a almost zero chance that any connection is recurring, entropy will approach infinity (the system is incompressible). To generalize this for our case, when entropy goes up, it means that communicative tokens interrelate in a more random way (i.e., the system is more

complex; i.e., has less compressible structure), while if entropy goes down, it means that communicative tokens show more structural interrelations.

Clustering. While entropy is a system-wide property, we can also study other relations between communicative tokens by assessing the degree to which they cluster or differentiate from each other. Clustering would indicate that there are multiple gestures that have similar features, which may indicate lack of differentiability. Indeed, we might expect that communicative tokens within a theme are likely to be ambiguous at beginning generations (e.g., the ambiguous reuse of the handcuffing gesture for 'to make an arrest' and 'police officer') and such gestures would cluster with edge weights of low D. It could then be that clustering becomes less over the generations as communicative tokens become maximally differentiable.

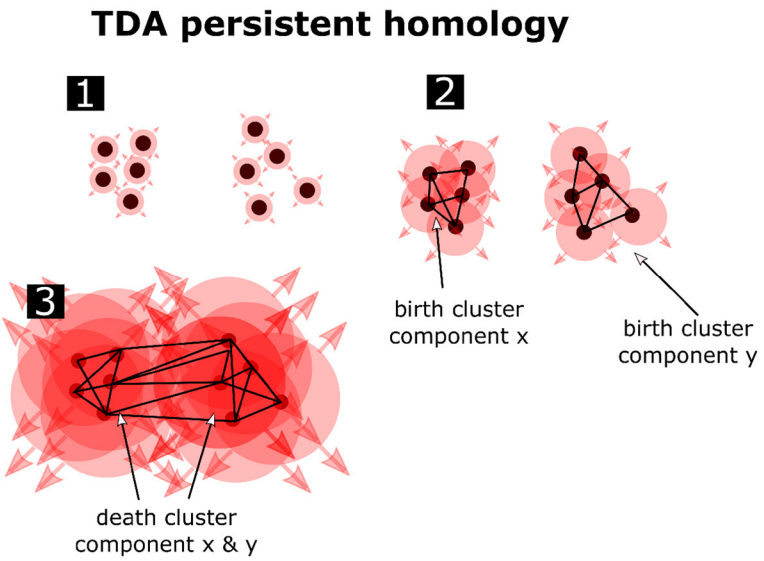
For the clustering measure we use a technique from topological data analysis (e.g., Sizemore, Phillips-Cremins, Ghrist, & Bassett, 2018) called persistent homology analysis (Bendich, Marron, Miller, Pieloch, & Skwerer, 2016; Otter, Porter, Tillmann, Grindrod, & Harrington, 2017), which can assess how stable (i.e., persistent) network components are through a continuous quantification.

Consider that the distance matrices contain coordinates for each gesture in a multidimensional space relative to all other gestures. Persistent homology measures the degree to which gestures cling together in a relatively stable fashion. Its measure can be visualized as involving gradually expanding circles around every gesture token (Figure 4). When circles touch, they merge to form a new cluster component. At the start of this process every single gesture is in its own 'cluster'. Soon enough circles begin to touch, forming new clusters of multiple gestures. Some such clusters will merge when their circles touch, others are so distant that they exist on their own for a longer while. When all circles have grown maximally, all nodes are connected and only a single overall cluster remains. Throughout this process, every cluster has its own lifetime (from emergence to assimilation). The average lifetime of reliable clusters is a measure of the amount of clustering in the network.

To compute cluster persistence, we used R-package 'TDAstats' (Wadhwa et al., 2019). We averaged persistence for the statistically significant components only, whereby we uses the 'TDAstats' own bootstrapping method (set at chance level of 0.975). The

selection of statistically reliable components was applied because many detected components are of very short persistence and reflect noise/chance level occurrences of components. We computed the average persistence of components (0-cycles) for each distance matrix (i.e., each individual's gesture network).

Figure 4. Network property example

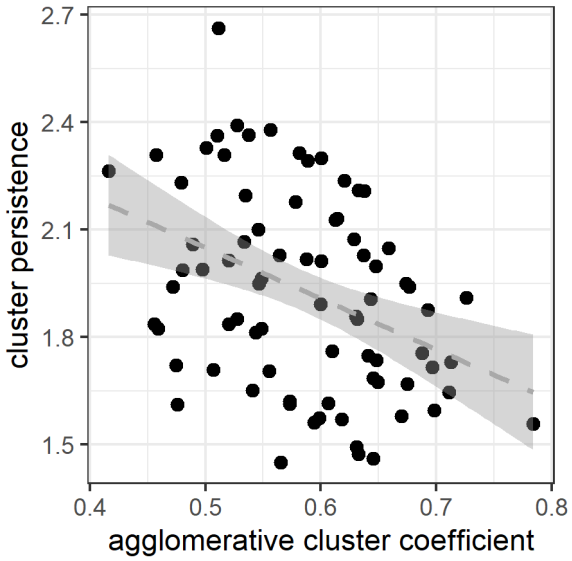


Note Figure 4. A visual example of the persistent homology procedure in Topological Data Analysis. Each token has a certain distance to all other tokens. Persistent homology analysis (PH) assesses the stability of components in this spatial organization by gradually increasing a spatial threshold (the radius) at which nodes get connected, indicated here by red growing radii. At 1, all tokens are unconnected. At 2, two distinct clusters x and y emerge. At 3, these two clusters merge into a single cluster z. The longer clusters survive at gradually increasing radii, the stabler they are.

Persistent homology is useful for multidimensional data structures like the weighted fully connected distance matrices we are working with. This is because it allows for a continuous quantification of cluster stability at multiple scales (clusters of clusters), in contrast to a binary assignment of nodes to particular clusters. Since Topological Data Analysis is relatively new analysis toolkit in cognitive science (Lum et al., 2013; Zhang, Kalies, Kelso, & Tognoli, 2020), we also made a comparison with another classic clustering measure: hierarchical clustering analysis with "average" linkage. For each matrix we computed the agglomerative clustering coefficient with R-package 'cluster', where a low clustering coefficient indicates more clustering in the data while a larger value indicates less clustering. When cluster persistence according to persistent

homology is high, the clustering coefficient is structurally lower (Figure 5), indicating that both measures converge on their estimate of 'clusteriness' of the data, r = -0.39, p < .001. Hereafter we only report cluster persistence as a measure of clusteriness.

Figure 5. Cluster measure comparison



Note Figure 5. Higher cluster persistence as measured by persistent homology is related to a lower agglomerative cluster coefficient in hierarchical cluster analysis, indicating that both measures are tracking a clustering property.

Kinematic Properties

We first selected five potential measures representative of kinematic quality of the movements in terms of segmentation, salience and temporality, namely submovements, intermittency, gesture space, rhythm, and temporal variability (or rhythmicity). See Figure 6 for two example time series from which most measures can be computed. All measures were computed for each keypoint' time series seperately and then averaged so as to get an overall score for the multimodal utterance as a whole. Based on these exploratory measures we eventually selected three measures tracking gesture segmentation (intermittency score), gesture salience (gesture space), gesture's temporality (temporal variability). Correlations and distributions are shown in Figure 7.

Gesture salience. As a measure for gesture salience or reduction, we computed a gesture space measure. This was determined by extracting the maximum vertical

amplitude of a keypoint multiplied by the maximum horizontal amplitude, i.e., the area in pixels that has been maximally covered by the movement.

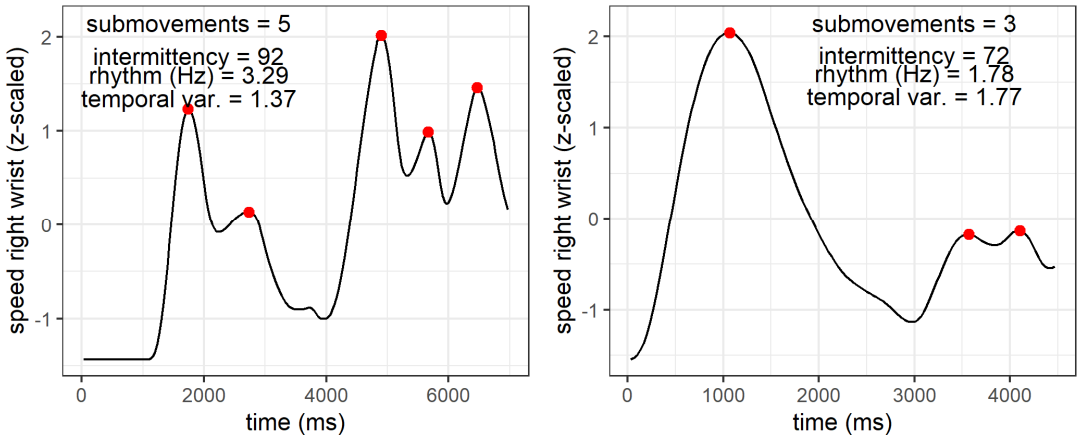
Gesture segmentation. We first computed a submovement measurement similarly implemented by Trujillo, Vaitonyte, Simanova, & Özyürek (2019). Submovements are computed with a basic peak finding function which identifies and counts maxima peaks in the movement speed time series. We set the minimum interpeak distance at 8 frames, and minimum height = -1 (z-scaled; 1 std.), minimum rise = 0.1 (z-scaled). We logtransformed the submovement measure due to a skewed distribution.

A property of the submovement measure is that it discretizes continuous information and uses arbitrary thresholds for what counts as a submovement, thereby risking information loss about subtle intermittencies in the movement. To have a more continuous measure of intermittency (the opposite of smoothness) of the movement we computed a dimensionless jerk measure (Hogan & Sternad, 2009). This measure is dimensionless in the sense that it is scaled by the maximum observed movement speed and duration of the movement. Dimensionless jerk is computed using the following formula $\int_{t2}^{t1} x'''(t)^2 dt) * \frac{D^3}{max(v^2)}$, where x''' is jerk, which is squared and integrated over time and multiplied by duration D cubed over the maximum squared velocity $max(v^2)$. As figure 6 shows, this measure correlates very highly with submovements, thus we chose to only use intermittency for further analysis. We logtransformed our smoothness measure due to a skewed distribution. Note that a *higher* intermittency score indicates more intermittent (less smooth) movement.

Gesture temporality. From the submovement measure we computed the average interval between each submovement (in Hz), which is a measure of rhythm tempo. This measure was, as expected, highly correlated with intermittency score, as tempo goes up when more segmented movements are performed in the same time window, $r = 0.61$, $p = < .001$, which led us to drop this measure for our analysis. Instead, we use another temporal measure that is more orthogonal to intermittency and gesture space, and which captures the stability of the rhythm, i.e., the temporal variability (the opposite of isochrony) of the movements. This measure is simply the standard deviation of the temporal interval between submovements (given in Hz): a higher score indicates more temporal variability and a lower score indicates more

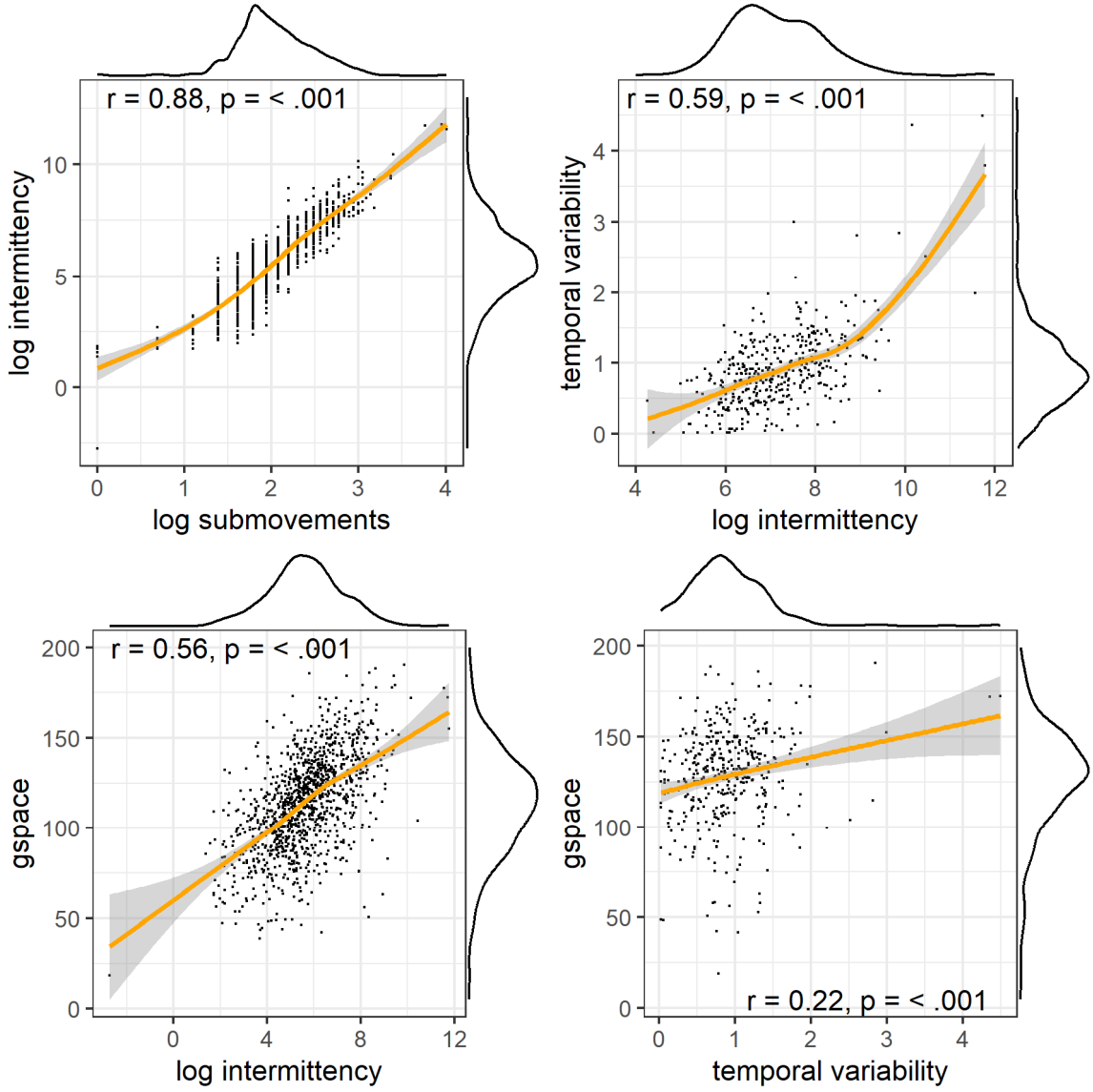
isochronous rhythm. Note, this measure cannot be calculated when there are less than 3 submovements (i.e., when there no intervals to detect the temporal variability of).

Figure 6. Overview kinematic measures



Note Figure 6. Two timeseries (belonging to two unique trials) are shown for right-hand wrist speed. From these time series, as well as the time series for other body parts, we computed measures tracking segmentation, namely, submovements (number of observed peaks in red) and intermittency. We further computed measures concerning temporality, namely the average time between submovements, i.e., rhythm in Hertz. We also computed temporal variability, which is the standard deviation of the rhythm in Hertz. Gesture space was calculated from the x,y position traces and is not shown here.

Figure 7. Correlations and distributions for kinematic measures per trial

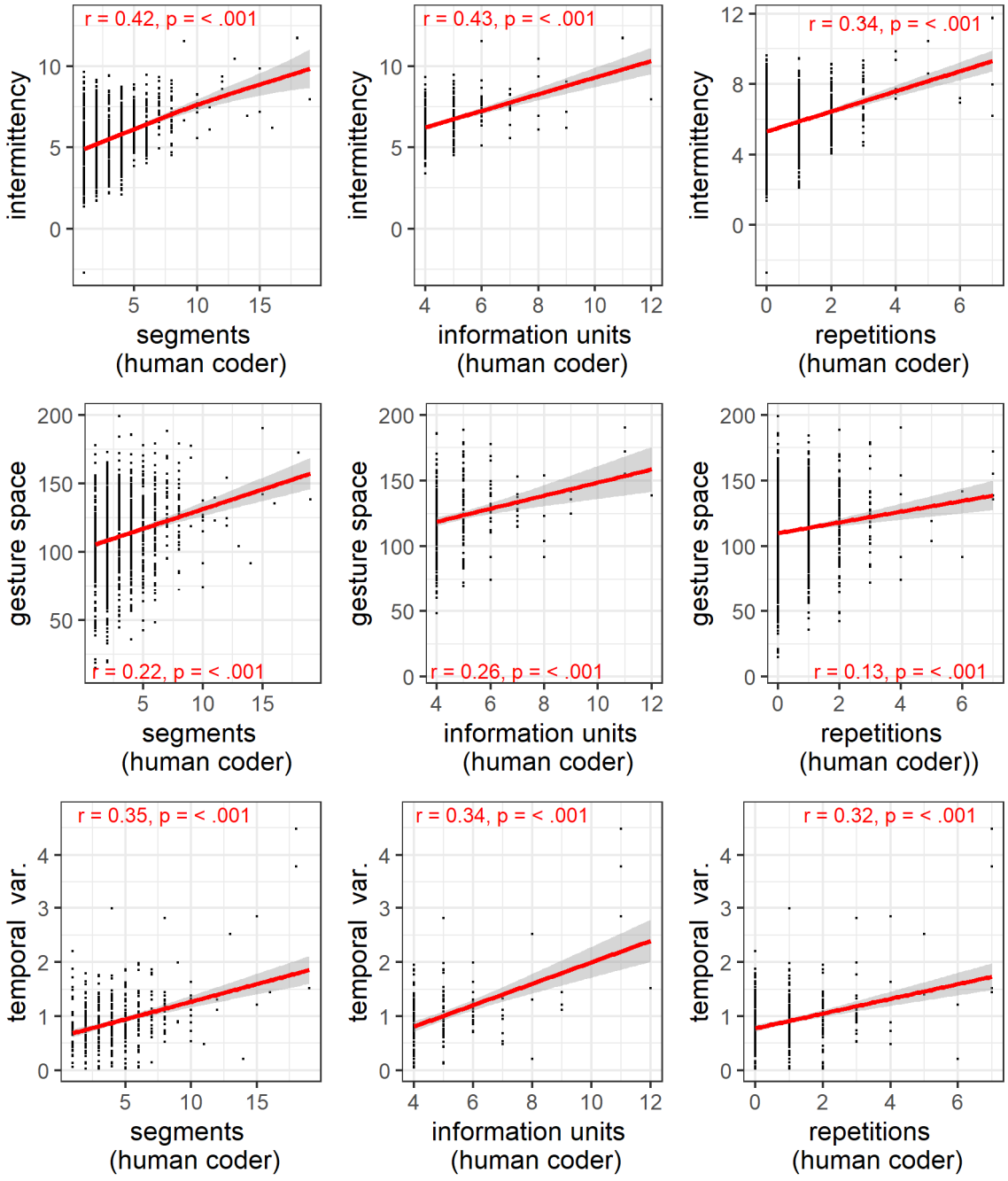


Note Figure 7. Left upper panel, correlations and distributions are shown for intermittency and submovement. Given their high correlation we will use intermittency score for our final analysis. Other correlations are shown for the selected measures, rhythmiticy, gesture space and intermittency.

Human coding and kinematic measures. It would be helpful to know how these automated kinematic measures approximate hand-coded data from Motamedi and colleagues (2019). The hand-coded data consisted of the amount of unique information units of the gesture utterance, the number of repetitions in the utterance, as well as the number of segments (information units + repetitions). We should predict that our kinematic intermittency score should correlate with the number of segments, repetitions

and information units as the kinematics will have to carry those information units by contrasts in the trajectories. Figure 8. shows the correlations for our kinematic measures and the human-coded gesture information. It shows that the amount of information units (unique, repeated or total) in the gesture as interpreted by a human coder are reliably correlating with kinematic intermittency (more intermittent more information), gesture space (larger space more information) and temporal variability (more stable rhythm more segments).

Figure 8. Correlations of kinematic measures with human-coded gesture information



Note Figure 8. On the horizontal axes the human-coded number of gesture segments, unique information units, and the number repetitions (of information units) are shown. On the vertical axes our automatic kinematic measures are shown: intermittency, gesture space, and temporal variability. The findings show that our measures are a proxy for human judgments, such that more intermittent kinematics reflect more information units (repeater and/or unique). Larger gesture space is related to more information units. Lower temporal variability in kinematics is further associated with more information units.

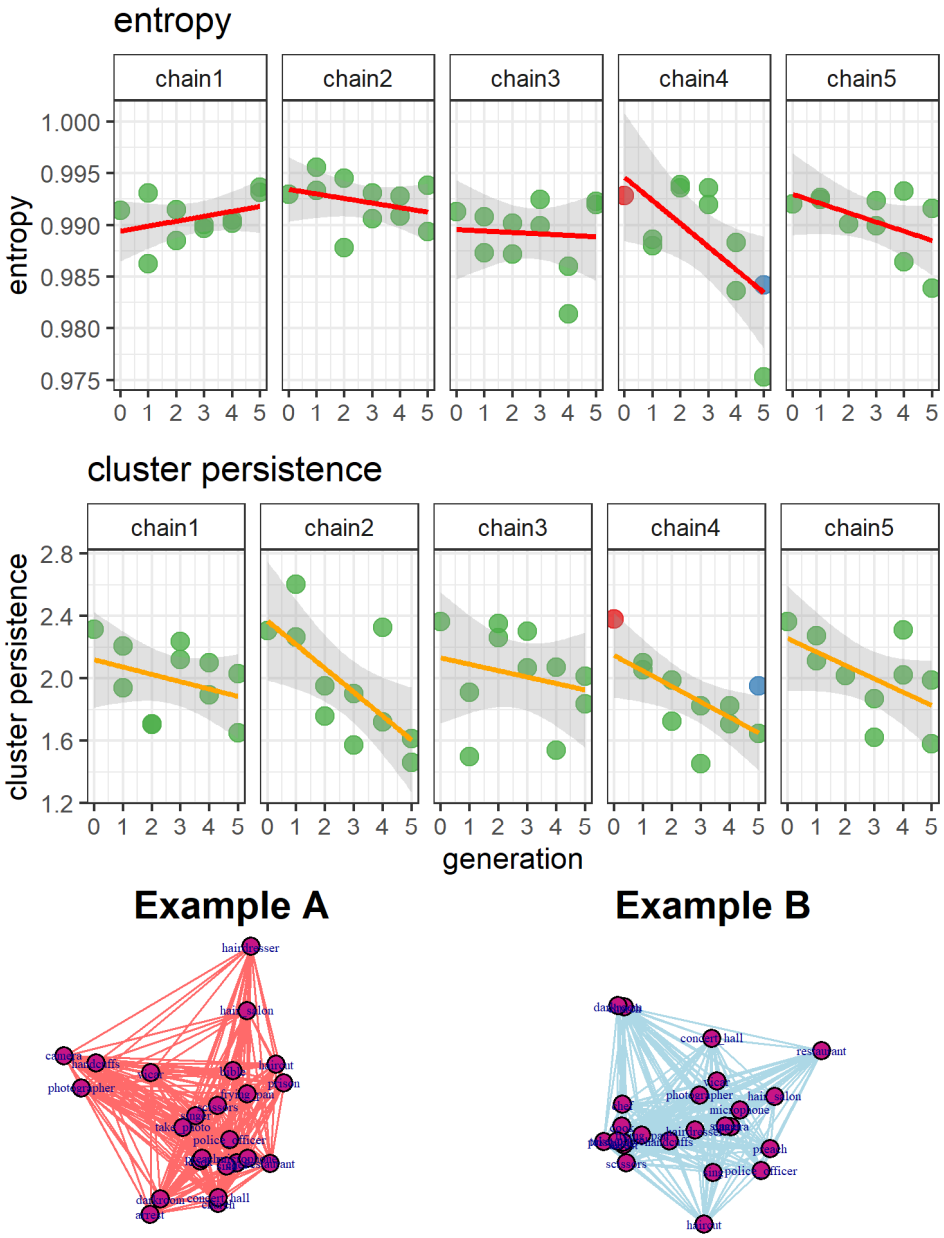
Main Results

We will first report findings on how relations between communicative tokens changed over the generations, as indicated by our network measures. We then validate whether network entropy approximates systematicity as observed by human coders. Subsequently, we will assess whether network changes occurred between particular tokens, namely the function vs. theme grouping. Finally, we will report on whether structural kinematic changes occurred over the generations for verb (action) and non-verb gestures (objects, persons, locations), and how such kinematic changes related to changes on the network level.

Network changes over generations

Figure 9 shows that for the gesture networks, that entropy was generally decreasing as a function of generation, indicating lower complexity of gesture interrelations as the system matures. Furthermore, there was less clustering at later generations (lower cluster persistence), indicating that kinematic patterns became more differientiable.

Figure 9. Changes in networks measures over generations within chains



Note Figure 9. For each chain the changes over generations in entropy and cluster persistence is shown, with generation 0 indicating the seed gesture set. For each generation > 0 there are two data points as there are two participants in each generation. Two example data points (red, and blue) are shown with their corresponding red and blue network representation (lower panel). In general cluster persistence decreased, indicating less differentiability between tokens. This may be seen in example A where there are relatively large cavities between tokens, while in example B the token organization is more homegonously tessellated. Indeed, entropy tends to decline over the generations, indicating that relationships between tokens became less diverse, possibly indicating systematicity in the way nodes are connected.

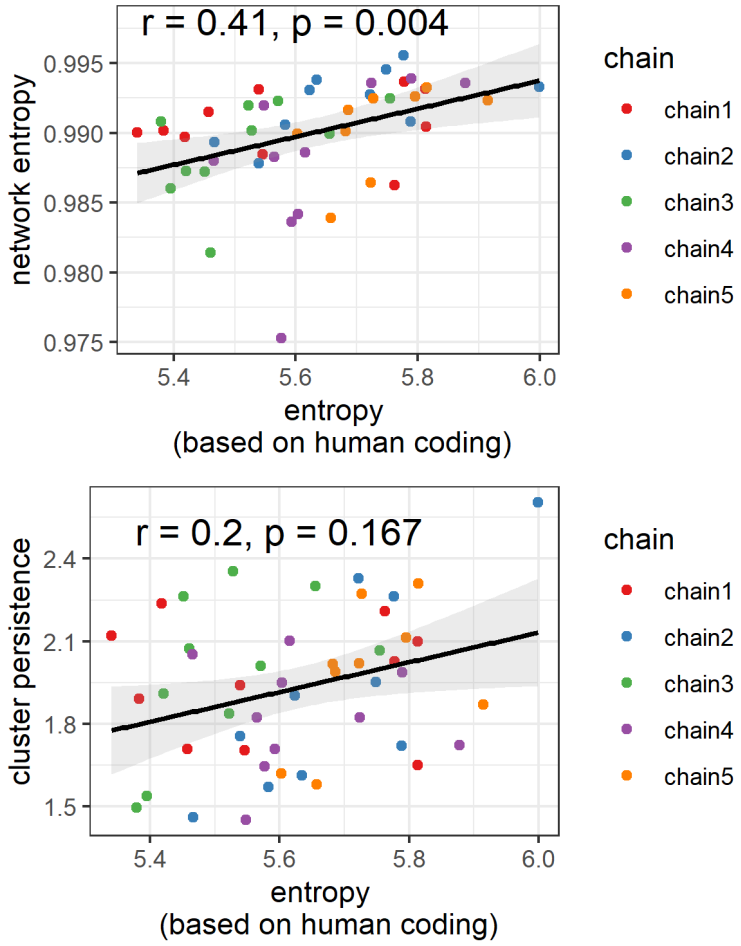
We tested these trends separately for each network property with mixed linear regression models, with chain as random intercept (random slopes did not converge for these models) and generation as independent predictor (0-5 generations, with generation 0 being the seed gesture network).

Generation was a reliable predictor for entropy as compared to a basemodel predicting the overall mean, chi-squared change (1) = 4.75, $p = 0.03$, model R-squared = 0.08. Model estimates showed that with increased generation the entropy decreased, b estimate = -0.0006, t (48.00) = -2.19, $p = 0.03$, Cohen's d = -0.63).

Cluster persistence was predicted by generation as compared to a basemodel, chi-squared change (1) = 14.60, $p <.001$, model R-squared = 0.24. Model estimates showed that with increased generation the cluster persistence decreased, b estimate = -0.09, t (48.00) = -4.02, $p < .001$, Cohen's d = -1.16). Note that the effect size of generation on cluster persistence is about twice as strong as compared to entropy.

We can further ask whether it is the case whether our network entropy measure is approximating the entropy of hand-coded gestures. Figure 10 confirms that this is indeed the case, such that entropy increase based on human-coded information units is related to increase in entropy based on gesture network entropy.

Figure 10. Gesture network entropy versus human-coded entropy



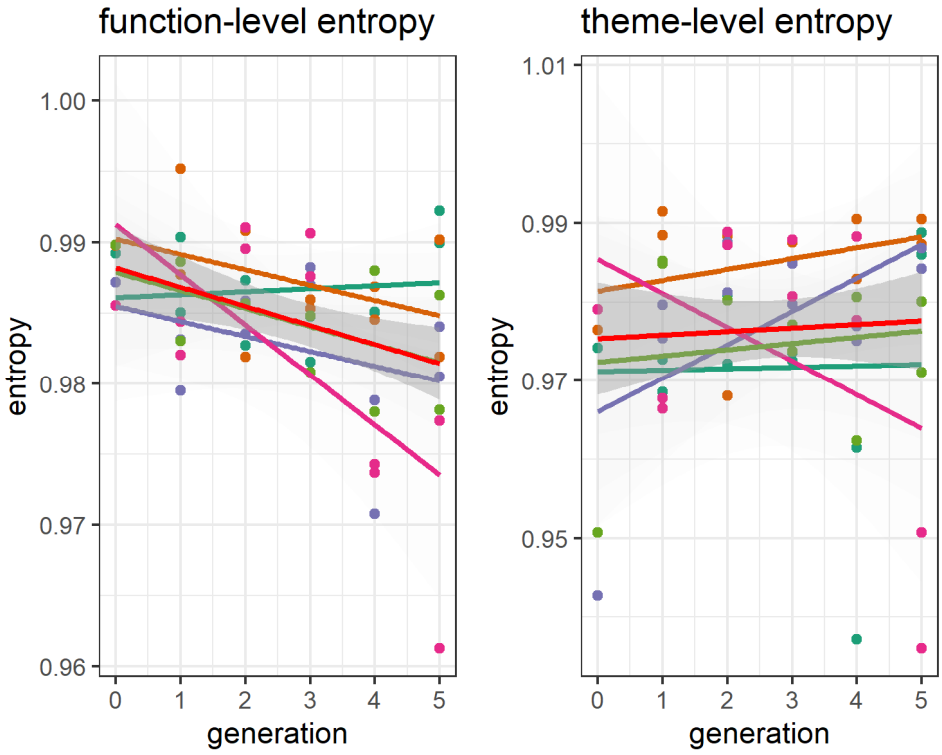
Note figure 10. The upper panel shows that there there is a strong relationship between the gesture network entropy with that of entropy computed on human-coded information units. That network entropy is uniquely related to systematicity is further corroborated by the finding that cluster persistence is not reliably correlated with entropy based on human coding. It does seem that gesture network analysis is a form-based proxy for systematicity in silent gesture.

Changes within theme versus changes within function

We can also localize where systematicity is most likely to increase (i.e., decrease in entropy) by subsetting the communicative tokens based on theme and function groupings. Note that theme marking gestures were not quantified in the original study, but functional marking gestures were and showed increase occurrences over the generations (Motamedi et al. 2019). For each participant we selected a sub-network grouped by function category gesture utterance or theme category gesture utterance and then computed network entropy for each of those subnetworks. This was done for all

category tokens (e.g., “action”, “agent”, etc.) and averaged for function and theme separately, to yield an average entropy for each category. See figure 11 for the main results of these subset networks.

Figure 11. Change in entropy in theme-level networks versus function-grouped networks



Note Figure 11. On the left panel, the average network entropy for the function-grouped gestures are plotted over the generations with red line showing the trend averaged over chain (other-colored lines). On the right panel this is shown for the gestures grouped by theme category. It can be seen that only the function-grouped gesture networks showed increased systematicity (lower entropy) over the generations.

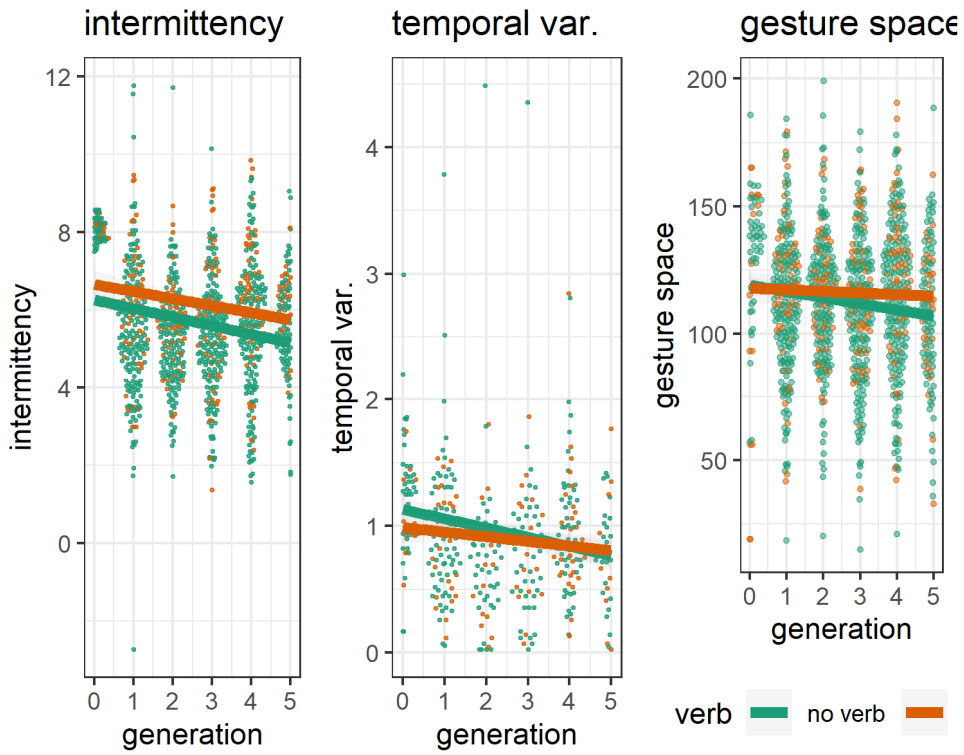
We find that only functionally grouped tokens were minimizing entropy over the generations. Including generations for predicting function-level network entropy increased predictability as compared to a base model (random intercept chain, random slopes did not converge), chi-squared change (1) = 8.99, $p < .001$, model R-squared = 0.15, with generation relating to lower entropy b estimate = -0.0014, t (48.00) = -3.08, $p < .001$, Cohen’s d = -0.89).

There was however no reliable decrease in entropy for the theme-level networks, chi-squared change (1) = 0.18, $p = 0.67$, model R-squared = 0.00.

Kinematic features

Next we performed mixed regression analysis for assessing potential kinematic changes as a function of generation, with random intercept for objects nested within chains (random slopes did not converge). See figure 12 for main results.

Figure 12. Change in kinematic properties over generations



Note Figure 12. Generation trends per chain are shown for intermittency, temporal variability and gesture space. Each observation indicates a communicative token, and these are spatially organized per their density distribution and colored by verb (green) or no verb (orange) (i.e., 'action' versus other function gestures). We can see that over the generations, movements become more smooth (lower intermittency score), with a more stable rhythm (lower temporal variability), and more minimized movements (smaller gesture space). Note, that temporal variability has lower data points as often the movement did not consist of more than 2 submovements. Thus, temporal variability indicates that *when there is a multi-segmented movement*, then such movements were more rhythmic.

Generations reliably predicted intermittency of the movements relative to a basemodel, chi-squared change (1) = 65.43, p <.001, model R-squared = 0.05. When adding verb as another predictor, this improved model fit for intermittency, chi-squared change (1) = 8.24, p < .001, model R-squared = 0.05. In this final model generation

predicted lower intermittency score, b estimate = -0.2109, t (1,098.00) = -8.19, $p <$ .001, Cohen's d = -0.49). Silent gestures conveying verbs showed lower intermittency in general, b estimate = 0.4448, t (1,098.00) = 2.95, $p <$ .001, Cohen's d = 0.18). There were no interaction effects of generation and verb.

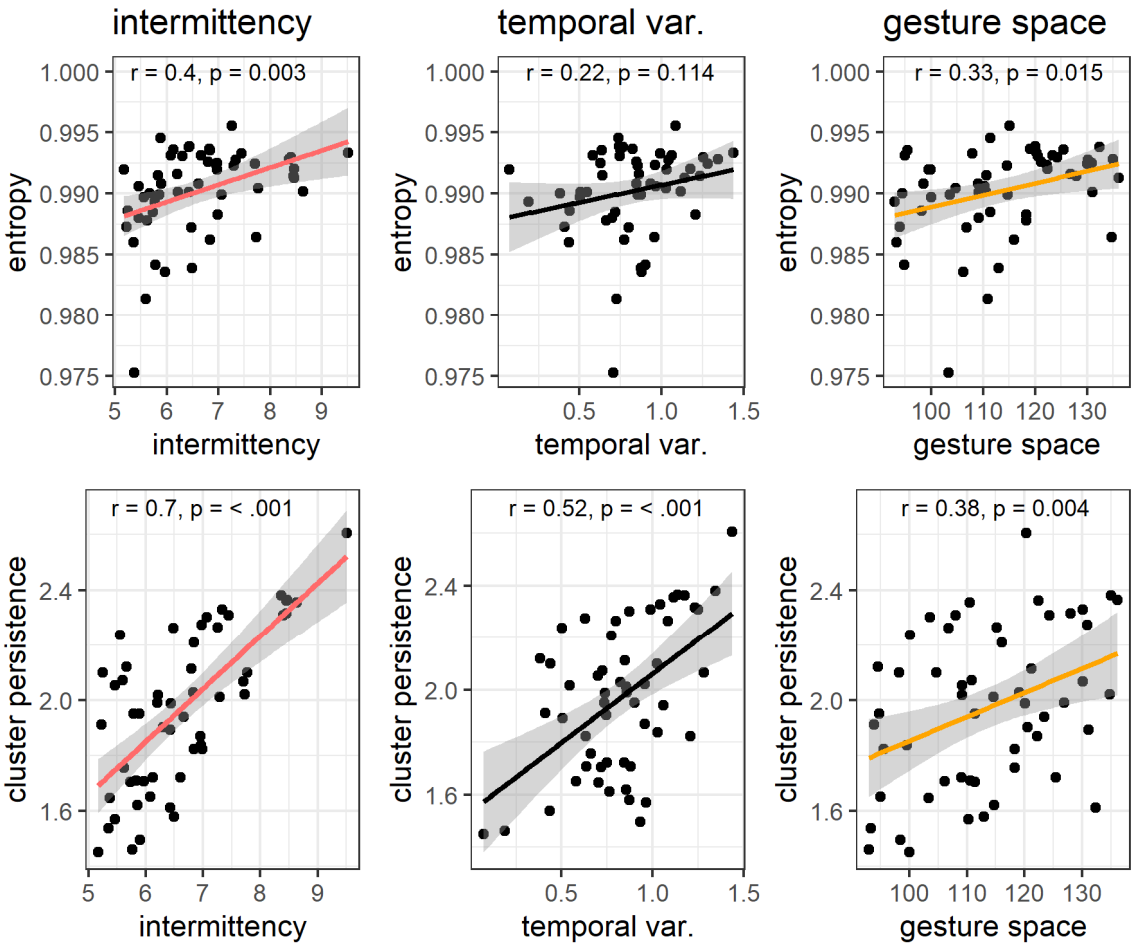
We also observe lower temporal variability as a function of generations, chi-squared change (1) = 21.03, $p <$.001, model R-squared = 0.04, indicating more stable rhythmic movements at later generations, b estimate = -0.0624, t (357.00) = -4.63, $p <$ .001, Cohen's d = -0.49. Adding verb or verb x generation to model temporal variability did not improve model fit. Finally, over the generations gesture space decreased, chi-squared change (1) = 19.86, $p <$.001. Model estimated gesture space was less for later generations, b estimate = -1.9968, t (1,130.00) = -4.47, $p <$ .001, Cohen's d = -0.27. Adding verb or verb x generation to model gesture space did not improve model fit.

In conclusion, our kinematic results show all the hallmarks of increased communicative efficiency. Namely, gestures were on average smaller, less temporally variable, and less intermittent as the communicative system matured. Silent gestures that conveyed a verb were generally less intermittent, suggesting that they consist of smoother movement patterns.

Relations between Kinematic and network properties

Figure 11 contains the correlations of the relationships of kinematic properties (average per participant) and the network measures cluster persistence and entropy. Network entropy goes down as the average gesture space decreases, and the movement becomes less intermittent. This also comes at a trade-off, such that this simplification of kinematics also reduces differientiability of communicative tokens as shown by less stable clustering when gesture become smaller, less temporally variable, and less intermittent. Thus on the kinematic level there seems to be a general decrease of complexity which is further reflected on the level of the system as a whole as utterances become less *kinematically* differientable (less clustering) and more structured in their relations (lower entropy).

Figure 13. Relation between kinematic properties and network measures



Note Figure 13. Correlations are shown for each kinematic property averaged over all utterances and the concomittant network measure result. It can be seen that less intermittency, lower temporal variability, and smaller gesture spaces, relate to lower entropy and lower cluster persistence. This indicates that complexity in movement is cashed out in terms of systematicity and more homegeneous interrelationships (lower clustering) on the network level.

Discussion

Based on signal processing alone we have detected systematic changes reflective of a linguistically maturing communication system from continuous multi-articulatory kinematics of silent gestures. We applied computer vision techniques to extract kinematics from video data, and then applied an analysis procedure to detect structural relations between gestural utterances (Pouw & Dixon, 2019). We found that communicative tokens showed higher systematicity at later generations, conceptually replicating results that were based on human coding of the gesture's content (Motamedi et al., 2019). Indeed, gesture network entropy turned out to be a good approximation of entropy based on human coding of the gesture content. We further find that tokens were less stably differentiable on the form level as tokens have lower cluster persistence over the generations. Moreover, we found a decrease in entropy for the functional rather than the thematic dimension. While in the original study no increase in efficiency was found based on measuring gesture information units, we did detect increases of communicative efficiency for gesture kinematics. Over generations, gestures became less segmented (more smoother), more rhythmic (if comprised of more than 3 submovements), and smaller. We also show that action gestures have a different kinematic quality as compared to non-action gestures, being more smooth (less intermittent) in their execution. Finally, we show that the decrease in kinematic complexity on the token level, predicts system-level changes of decreased entropy and decrease in clustering.

That entropy decreased for gestures within the functional category at the level of kinematics, is consonant with the human coding findings of the original study and other related findings on sign languages showing regular employment of functional categories such as object- versus action distinctions (Padden et al., 2013). That gestures referring to actions are less intermittent as compared to non-action gestures in terms of their kinematics, conceptually replicates research based on human coding showing that action gestures often consist of a single segment (Ortega & Özyürek, 2020b). In sum, our findings indicate that kinematics are revealing of the functional nature of gesture references, showing unique trajectories of change during iterated learning.

A decrease in cluster persistence over generations here is likely to reflect the differentiability of communicative tokens, which as originally reported often showed iconic gestures at early stages in the iterations that were sometimes ambiguous in the

theme category, and maximally differentiated from the other-themed gestures. For example, "arrest" and "police officer" could both contain a gesture that enacts the appliance of hand cuffs. Thus within themes there was clustering, but across themes there is differentiation. When gestures are disambiguated over the generations this will result in increased distances among the gestures within this category on the network level, i.e., leads to less clustering. While clusters became more unstable over the generations, the diversity of the interrelationships of the communicative tokens decreased (i.e., entropy decreased), and this is especially on the functional level. This suggests that there is a more consistent and thus homogeneous way in which the communicative system is organized, and the reorganization is caused by a reuse of gesture *across themes* and *within function*. That this increase in consistency is indeed a form of systematicity is fully supported by the detection of entropy decrease over the generations for communicative tokens grouped on the functional dimension (e.g., agent, action, location), but not the thematic dimension (e.g., justice, cooking).

The kinematic findings suggest that the manual utterances simplify, in the way of reducing in size, in the reduction of submovements, and the decrease in temporal variability of the utterance if it comprised of multiple submovements. This simplification seems to be a reduction in articulatory effort, as making a minimal amount of smaller rhythmic movements reduces the degrees of freedom for articulation (Bernstein, 1967; Kelso et al., 1983). Moreover, this increased rhythmicity could also increase learnability and comprehensibility of the gesture, as we know from speech perception in noisy conditions that it is optimally perceived when speech is more rhythmic (Wang, Kong, Zhang, Wu, & Li, 2018).

Interestingly, this reduction of degrees of freedom of the pronunciation, is precisely what one finds for novice learners of ASL. ASL learners have been found to spatially reduce their signs as they become more fluent (Lupton & Zelaznik, 1990; Wilbur, 1990). Moreover, a reduction in duration between the compounds of the signs have been observed during learning progression, where multicomponent component signs are increasingly performed as a single sign. In the present paradigm, there is a similar evolution of pronunciation, such that gestural multi-articulatory utterances acquire stable functional organizations across generations. Suboptimal organization of sub-movements will be filtered out as it were over the generations, and the temporally

extended movement sequence becomes likely more coordinated whereby degrees of freedom are reduced by functioning as a single multi-aticulatory coordinative structure (Bernstein, 1967; Kelso, et al., 1983), affecting for example gesture's temporal variability and intermittency. That head movements improved differientation of real vs. falsely paired gestures in our analysis, further emphasizes that multiple articulators coordinated in the production of meaning in the current task. This finding resonates with the known grammatic, phonetic, and prosodic functions that head movements have in sign languages such as ASL (Tyrone & Mauk, 2016). Indeed, as Sandler (2018) has argued for sign languages, the expressive power of the body lies in the combination of different articulators which can attain unique linguistic functions which can then be combined in parallel into a single linguistically complex utterance.

Note that our method allowed to account for the multiarticulatory nature of communication without formalized additional coding of the head movements, and we were able to quantify the unique communicative contributions of head and upper limb movements in the current paradigm. In this way, the current method is a bottom-up approach that will invite further investigation when needed. Our bottom-up approach further showed that gesture network entropy decreases alongside the entropy obtained from human coded content segmentation of the gestural utterance (Motamedi et al., 2019), suggesting that systematicity in form can be detected without the need for an apriori coding scheme.

But our method as exposed here goes one step further. If gesture network analysis is complemented with kinematic feature analysis, it can be further assessed *what* is driving systematicity, providing insights on the evolution of the morphology of the silent gesture system. Coding schemes are notoriously difficult to formalize as any gesture researcher will confirm, and the current buttom-up method provides a formalized procedure for the detection of gesture evolution. As it is formalized and reproducible, the method is waiting to be applied to large datasets that are impractical to (completely) code by human annotators. An exciting avenue of further research is how different morphological evolutions can yield similar or different levels of systematicity at the gesture network level depending on different communicative constraints in vast populations. As such, we have shown that human-coded information units can be approximated from the kinematics (intermittency) and it can be assessed whether

changes in such units are cashed out on the system level in the form of differientiability (clustering) and information compression (network entropy) - no human coding needed.

There are two important caveats to the analyses presented here. First, in general, it is the case that kinematic analysis cannot say anything about the precise semiotic content that might evolve, and this is especially the case with incereasing 'drifts towards the arbitrary' (Tomasello, 2008). Although such drift might be detected via our network analysis, recognizing its possible semiotic content will always require extensive analysis (Sandler, 2018). However, since changes can be detected our analysis may invite human coding to be performed on a subset of the data which show promising changes over generations, inviting further necessary human interperation to understand what was driving such changes on the level of meaning.

A second caveat is that there is a limitation to dynamic time warping. If we appreciate that compositionality increases as a communicative system matures, holistic gestures become segmented and the order of presentation of such segments might be varied. However, the dynamic time warping algorithm is sensitive to ordering and would judge two gestures containing identical segments in different orders as very different, while for a human coder the similarity between differently ordered segments might be transparent. Thus our analysis may at times judge sequences of gestures highly dissimilar when in fact they are merely ordered differently. There are ways to circumvent this by only look for trajectory overlaps rather than ordering through time (Pouw & Dixon 2019), but such analysis goes beyond the current approach.

Both of these caveats mean that our approach to kinematics, like all quantitative analyses of human behavior, requires some degree of human oversight (for meaningful implementation) and human insight (for judicious interpretation). When these requirements are met, we believe that our fully reproducible and automatable methods can make important contributions: It will reduce the amount of manual coding which is currently consuming many researchers' time. It provides a much needed *multiscale* approach to how gestures evolve as communicative systems. The current method can further scale up the study of language evolution across modalities, as the kinematic analysis shown here functions much like an acoustic and articulatory analysis in speech. Beyond such promises, the current multiscale approach has shown that silent gesture kinematics evolve in structural ways during iterated learning.

References

Aronoff, M., Meir, I., Padden, C. A., & Sandler, W. (2008). The roots of linguistic organization in a new language. *Interaction Studies*, *9*(1), 133–153. doi:10.1075/is.9.1.10aro

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, *58*(2), 495–520. doi:10.1016/j.jml.2007.02.004

Bendich, P., Marron, J. S., Miller, E., Pieloch, A., & Skwerer, S. (2016). Persistent Homology Analysis of Brain Artery Trees. *The Annals of Applied Statistics*, *10*(1), 198–218. doi:10.1214/15-AOAS886

Bernstein, N. (1967). *The Co-ordination and Regulations of Movements* ([1st English ed.] edition.). Pergamon Press.

Bickerton, D. (2009). *Adam's Tongue*. New York: Hill & Wang.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1302–1310). Honolulu, HI: IEEE. doi:10.1109/CVPR.2017.143

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*. doi:10.1017/S0140525X1500031X

Claidière, N., Smith, K., Kirby, S., & Fagot, J. (2014). Cultural evolution of systematically structured behaviour in a non-human primate. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1797), 20141541. doi:10.1098/rspb.2014.1541

Corballis, M. C. (2002). *From hand to mouth: The origins of language*. Princeton, NJ.: Princeton University Press.

Cornish, H., Dale, R., Kirby, S., & Christiansen, M. H. (2017). Sequence Memory Constraints give rise to language-like structure through iterated learning. *PLoS ONE*, *12*(1), e0168532. doi:10.1371/journal.pone.0168532

Csárdi, G. (2019). Package 'igraph' network analysis and bisualization (Version 1.2.4.1). Retrieved from http://bioconductor.statistik.tu-dortmund.de/cran/web/packages/igraph/igraph.pdf

Deutscher, G. (2005). *The unfolding of language: An evolutionary of mankind's greatest Invention*. New York: Metropolitan Books.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*(10), 603–615. doi:10.1016/j.tics.2015.07.013

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., ... Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PLoS ONE*, *10*(9). doi:10.1371/journal.pone.0136100

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Boston: Harvard University Press.

Enfield, N. J. (2016). *Natural causes of language: Frames, biases, and cultural transmission*. Berlin: Language Science Press. Retrieved from https://langsci-press.org/catalog/book/48

Engesser, S., & Townsend, S. W. (2019). Combinatoriality in the vocal systems of nonhuman animals. *WIREs Cognitive Science*, *10*(4), e1493. doi:10.1002/wcs.1493

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*(3), 351–386. doi:10.1111/j.1551-6709.2009.01090.x

Garrod, S., Fay, N., Lee, J., Oberlander, J., & Macleod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, *31*(6), 961–987. doi:10.1080/03640210701703659

Gerwing, J., & Bavelas, J. (2004). Linguistic influences on gesture's form. *Gesture*, *4*(2), 157–195. doi:10.1075/gest.4.2.04ger

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407. doi:10.1016/j.tics.2019.02.003

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in *R* : The **Dtw** package. *Journal of Statistical Software*, *31*(7). doi:10.18637/jss.v031.i07

Goldin-Meadow, S., So, W. C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, *105*(27), 9163–9168. doi:10.1073/pnas.0710060105

Haviland, J. B. (2013). The emerging grammar of nouns in a first generation sign language: Specification, iconicity, and syntax. *Gesture*, *13*(3), 309–353. doi:10.1075/gest.13.3.04hav

Hogan, N., & Sternad, D. (2009). Sensitivity of smoothness measures to movement duration, amplitude and arrests. *Journal of Motor Behavior*, *41*(6), 529–534. doi:10.3200/35-09-004-RC

Holler, J., & Wilkin, K. (2011). An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. *Journal of Pragmatics*, *43*(14), 3522–3536. doi:10.1016/j.pragma.2011.08.002

Kelso, J. A. S., Tuller, B., & Harris, K. (1983). A "dynamic pattern" perspective on the control and coordination of movement. In *The production of speech*. Berlin: Springer-Verlag. Retrieved from https://link.springer.com/chapter/10.1007/978-1-4613-8202-7_7

Kendon, A. (2017). Reflections on the "gesture-first" hypothesis of language origins. *Psychonomic Bulletin & Review*, *24*(1), 163–170. doi:10.3758/s13423-016-1117-3

Kirby, S., & Christiansen, M. H. (2003). From language learning to language evolution. In M. H. Christiansen & S. Kirby (Eds.), *Language Evolution* (pp. 272–294). Oxford University Press. doi:10.1093/acprof:oso/9780199244843.003.0015

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human

language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686. doi:10.1073/pnas.0707835105

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114. doi:10.1016/j.conb.2014.07.014

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651). doi:10.1098/rstb.2013.0302

Lou-Magnuson, M., & Onnis, L. (2018). Social network limits language complexity. *Cognitive Science*, *42*(8), 2790–2817. doi:10.1111/cogs.12683

Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., ... Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. *Scientific Reports*, *3*. doi:10.1038/srep01236

Lupton, L. K., & Zelaznik, H. N. (1990). Motor learning in sign language students. *Sign Language Studies*, *1067*(1), 153–174. doi:10.1353/sls.1990.0020

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, *5*(1). doi:10.1371/journal.pone.0008559

McNeilage, P. (2008). *The origin of speech*. New York: Oxford University Press.

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, *192*, 103964. doi:10.1016/j.cognition.2019.05.001

Mueen, A. K., & Keogh, E. (2016). Extracting optimal performance from dynamic time warping. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2129–2130). doi:10.1145/2939672.2945383

Muller, M. (2007). *Information retrieval for music and motion*. Heidelberg, Germany: Springer.

Namboodiripad, S., Lenzen, D., Lepic, R., & Verhoef, T. (2016). Measuring conventionalization in the manual modality. *Journal of Language Evolution*, *1*(2), 109–118. doi:10.1093/jole/lzw005

Ortega, G., & Özyürek, A. (2020a). Systematic mappings between semantic categories and types of iconic representations in the manual modality: A normed database of silent gesture. *Behavior Research Methods*, *52*(1), 51–67. doi:10.3758/s13428-019-01204-6

Ortega, G., & Özyürek, A. (2020b). Types of iconicity and combinatorial strategies distinguish semantic categories in silent gesture across cultures. *Language and Cognition*, *12*(1), 84–113. doi:10.1017/langcog.2019.28

Ortega, G., Schiefner, A., & Ozyurek, A. (2019). Hearing non-signers use their gestures to predict iconic form-meaning mappings at first exposure to signs - ScienceDirect. *Cognition*, *191*(103996). doi:10.1016/j.cognition.2019.06.008

Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., & Harrington, H. A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Science*, *6*(1), 17. doi:10.1140/epjds/s13688-017-0109-5

Padden, C. A., Meir, I., Hwang, S.-O., Lepic, R., Seegers, S., & Sampson, T. (2013). Patterned iconicity in sign language lexicons. *Gesture*, *13*(3), 287–308. doi:10.1075/gest.13.3.03pad

Pouw, W., & Dixon, J. A. (2019). Gesture networks: Introducing dynamic time warping and network analysis for the kinematic study of gesture ensembles. *Discourse Processes*. doi:10.1080/0163853X.2019.1678967

Pouw, W., & Trujillo, J. P. (2019). *Materials Tutorial Gespin2019 - Using video-based motion tracking to quantify speech-gesture synchrony*. Retrieved from 10.17605/OSF.IO/RXB8J

Ravignani, A., Delgado, T., & Kirby, S. (2016). Musical evolution in the lab exhibits rhythmic universals. *Nature Human Behaviour*, *1*(1, 1), 1–7. doi:10.1038/s41562-016-0007

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1907), 20191262. doi:10.1098/rspb.2019.1262

Sandler, W. (2018). The body as evidence for the nature of language. *Frontiers in Psychology*, *9*. doi:10.3389/fpsyg.2018.01782

Sato, A., Schouwstra, M., Flaherty, M., & Kirby, S. (2020). Do all aspects of learning benefit from iconicity? Evidence from motion capture. *Language and Cognition*, *12*(1), 36–55. doi:10.1017/langcog.2019.37

Schouwstra, M. (2017). Temporal structure in emerging language: From natural data to silent gesture. *Cognitive Science*, *41*(S4), 928–940. doi:10.1111/cogs.12441

Scott-phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, *14*, 411–417. doi:10.1016/j.tics.2010.06.006

Senghas, A., Kita, S., & Özyürek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in nicaragua. *Science*, *305*(5691), 1779–1782. doi:10.1126/science.1100199

Sizemore, A. E., Phillips-Cremins, J., Ghrist, R., & Bassett, D. S. (2018). The importance of the whole: Topological data analysis for the network neuroscientist. Retrieved from http://arxiv.org/abs/1806.05167

Slonimska, A., Özyürek, A., & Capirci, O. (2020). The role of iconicity and simultaneity for efficient communication: The case of Italian Sign Language (LIS). *Cognition*, *200*, 104246. doi:10.1016/j.cognition.2020.104246

Tomasello, M. (2008). *The origins of human communication*. Cambdride, MA: MIT press.

Trujillo, J. P., Vaitonyte, J., Simanova, I., & Özyürek, A. (2019). Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research. *Behavior Research Methods*, *51*(2), 769–777. doi:10.3758/s13428-018-1086-8

Tyrone, M. E., & Mauk, C. E. (2016). The Phonetics of Head and Body Movement in the Realization of American Sign Language Signs. *Phonetica*, *73*(2), 120–140. doi:10.1159/000443836

Verhoef, T., Kirby, S., & de Boer, B. (2016). Iconicity and the emergence of combinatorial structure in language. *Cognitive Science*, *40*(8), 1969–1994. doi:10.1111/cogs.12326

Wadhwa, R., Dhawan, A., Williamson, D., Scott, J., Brunson, J. C., & Ochi, S. (2019). TDAstats: Pipeline for topological data analysis (Version 0.4.1). Retrieved from https://CRAN.R-project.org/package=TDAstats

Wang, M., Kong, L., Zhang, C., Wu, X., & Li, L. (2018). Speaking rhythmically improves speech recognition under "cocktail-party" conditions. *The Journal of the Acoustical Society of America*, *143*(4), EL255. doi:10.1121/1.5030518

Wilbur, R. B. (1990). An experimental investigation of stressed sign production. *International Journal of Sign Linguistics*, *1*(1).

Zhang, M., Kalies, W. D., Kelso, J. A. S., & Tognoli, E. (2020). Topological portraits of multiscale coordination dynamics. *Journal of Neuroscience Methods*, *339*, 108672. doi:10.1016/j.jneumeth.2020.108672