Article:

ORIGINAL MANUSCRIPT

Many morphs: Parsing gesture signals from the noise

Alexander Mielke[1,2] · Gal Badihi[1] · Kirsty E. Graham[1] · Charlotte Grund[1] · Chie Hashimoto[3] · Alex K. Piel[4,5] ·
Alexandra Safryghin[1] · Katie E. Slocombe[6] · Fiona Stewart[4,5] · Claudia Wilke[6] · Klaus Zuberbühler[7] ·
Catherine Hobaiter[1]

© The Author(s) 2024

Abstract

Parsing signals from noise is a general problem for signallers and recipients, and for researchers studying communicative systems. Substantial efforts have been invested in comparing how other species encode information and meaning, and how signalling is structured. However, research depends on identifying and discriminating signals that represent meaningful units of analysis. Early approaches to defining signal repertoires applied top-down approaches, classifying cases into predefined signal types. Recently, more labour-intensive methods have taken a bottom-up approach describing detailed features of each signal and clustering cases based on patterns of similarity in multi-dimensional feature-space that were previously undetectable. Nevertheless, it remains essential to assess whether the resulting repertoires are composed of relevant units from the perspective of the species using them, and redefining repertoires when additional data become available. In this paper we provide a framework that takes data from the largest set of wild chimpanzee (*Pan troglodytes*) gestures currently available, splitting gesture types at a fine scale based on modifying features of gesture expression using latent class analysis (a model-based cluster detection algorithm for categorical variables), and then determining whether this splitting process reduces uncertainty about the goal or community of the gesture. Our method allows different features of interest to be incorporated into the splitting process, providing substantial future flexibility across, for example, species, populations, and levels of signal granularity. Doing so, we provide a powerful tool allowing researchers interested in gestural communication to establish repertoires of relevant units for subsequent analyses within and between systems of communication.

✉ Alexander Mielke
mielke.alexand@gmail.com

1 Wild Minds Lab, School of Psychology and Neuroscience, University of St Andrews, St Andrews, UK

2 School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK

3 Primate Research Institute, Kyoto University, Kyoto, Japan

4 Department of Anthropology, University College London, London, UK

5 Department of Human Origins, Max Planck Institute of Evolutionary Anthropology, Leipzig, Germany

6 Department of Psychology, University of York, York, UK

7 Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland

Introduction

If you want to see a biologist struggle, ask them to define what a species is. None of the easy answers apply to the real world. We spend a substantial amount of our time trying to fit messy, highly variable data into tidy, artificial boxes. In practice, we know that perfectionism is largely futile and accept that, at best, we are describing a fraction of the variation we see expressed in the world around us. Nevertheless, our work is often founded on comparison—we fit data into groups to be able to interrogate them. We classify individuals into species, populations, sexes, age groups. We classify behaviour using carefully constructed ethograms (Bateson & Martin, 2021). In practice the expression of all of these is graded (Hey et al., 2003), and we can choose different levels of detail—lumping or splitting sets—depending on our question of interest (or on the data available).

A clear example of this process is the study of communication, both in other species and in our own. First, we assign signals to distinct channels such as vocalisations and gestures, despite use of these channels overlapping extensively (Slocombe et al., 2011; Kendon, 1980). We then take the rich, graded systems expressed across those channels and carve them up into sets or 'repertoires' based on our intuition and ability to perceive and categorise streams of graded or blended signals into discrete units (Parr et al., 2005; Hobaiter & Byrne, 2017; Stokoe 1960). This parsing process is something the human brain is particularly good at (Saffran et al., 1996)—spoken human language as well as language-accompanying gestures and facial signals are produced not as a set of discrete signals, but as a continuous stream of information that is perceived discretely (Jack et al., 2014; McNeil, 1992). Doing so during early language learning is automatic: we are rarely aware that we are rapidly processing an extremely rich set of statistical cues (Saffran et al., 1999; Swingley, 2005).

When we rely on this intuitive classification to ask questions about communication—about information and meaning, structure and syntax, dialect and accent, individual variation and community or species-wide shared characteristics—our results become conditional on the choices we make in categorising signals, opening the door for considerable biases due to researcher degrees of freedom (Wicherts et al., 2016). Historically, in non-human primate (hereafter primate) vocal and facial communication, both highly graded systems, we often started by deriving minimal sets of stereotypical signals (e.g., 'hoo') that were then further qualified by context ('resting hoo', 'alarm hoo', 'travelling hoo'; e.g., Marler, 1976). (Interestingly, early work on human co-speech gesture took a similar approach: a treatise on communication for clergymen discriminated gestures of the same type into different contexts or 'styles' (epic, colloquial; Austin 1806).) Doing so typically involved discarding significant portions of data where signals fell outside of classification boundaries—for example, chimpanzee (*Pan troglodytes*) 'neutral' facial signals contain a lot of subtle muscle activation (Parr & Waller, 2006). As a result, we may have underestimated the actual amount of information being transmitted. For example, the 'silent bared teeth' display in crested macaques (*Macaca nigra*) was recently shown to comprise at least four distinct forms that varied reliably with context (Clark et al., 2020). Similarly to 'silent bared teeth' faces, early descriptions of unified call types have been slowly split into more detailed repertoires based on acoustic features and differences in use (e.g., Clay et al., 2015; Crockford et al., 2018; Slocombe & Zuberbühler, 2005, 2007). Advances in detailed acoustic analysis introduced a more data driven and replicable approach to studies of vocal communication (Sainburg et al., 2020). Critically, only in some cases have subsequent playback experiments confirmed that listeners meaningfully distinguish between these call variants (Slocombe et al., 2009; Slocombe et al., 2010). For many species, particularly those with highly graded call systems, the exact number of call types in the repertoire has remained elusive, with variation increasingly described using fuzzy element boundaries and clustering algorithms (Fischer et al., 2017; Wadewitz et al., 2015).

Researchers traditionally approached primate gestures differently, in that the system was historically split into many more distinct units than the vocal or facial channels, with little additional detail about signal production, so signals appeared less graded and more like a rich repertoire of easily differentiated units or 'gesture types' (Plooij, 1979; Tomasello et al., 1985). The difference between discrete gesture types has usually been established based on a mixture of morphological and contextual definitions designed to achieve high reliability within studies, but less consistency between studies (Rodrigues et al., 2021). This process also potentially splits gesture units along arbitrary lines where features salient to a human coder may not be detected or relevant to non-human primate individuals using them. While many of the challenges faced by coders of non-human primate gestures are shared by those studying human co-speech and co-sign gesture—and early researchers noted substantial similarities of gesture expression in young chimpanzees and children (e.g., Ladygina-Kohts & Vekker, 2002)—our similarities in body plan can lead to potentially misleading anthropocentric assumptions. For example, partner contact with fingers ('poke'), palm ('slap'), or fist ('punch') are associated with overlapping patterns of goals in chimpanzees (Hobaiter & Byrne, 2017), despite sometimes being coded as discrete 'types' by human observers (Hobaiter & Byrne, 2011a, b). Conversely, leaf-clipping is often lumped by coders, but when we split them according to features such as the location of the tearing action on the leaf, different variants become apparent (Badihi et al., 2023). More recently, there has been a push for replicable definitions of gesture types, and a focus on bottom-up detailed coding to allow for more granular analyses of gestural communication. For example, the use of 'minimum action units' or MAU (Grund et al., 2023). Here, as in studies of human co-speech gesture (Kendon, 2004), the initial physical movement of the gesture (equivalent to the preparation and action stroke; Kendon, 2004) is distinguished from the hold or repetition phases. Original coding is highly split with systematic detailed descriptions of gesture form, i.e., MAU, alongside gesture modifiers (Grund et al., 2023). We call these additional morphological characteristics of a gesture action 'modifiers' (termed 'features' in previous publications, e.g., Hobaiter & Byrne, 2017), as they modify the appearance of the gesture action. (Note: we recognise that terms such as gesture, action, and modifier are used heterogeneously across studies of human and non-human communication, as well as in other fields. For clarity, we provide a short set of definitions specifying our usage below, Table 1; a full expansion of these

Table 1 Gesture terminology

Term	Definition
Gesture action	The bodily movement that describes the current gesture instance. Note that all cases of gesture use, by our definition, must be produced as intentional signals.
Modifier	Additional characteristics of the production of a gesture action in a specific instance of use. These include, amongst others, the signaller's body part (or articulator), laterality, object use, and presence of flexion (at fingers, wrist, elbow).
Gesture morph	Gesture forms that represent non-random combinations of gesture actions and a particular set of modifiers. Morphs can be constructed at multiple levels of splitting, depending on the number and type of modifiers included.
Minimum action unit (MAU)	The part of the gesture action that contains the minimal information needed to discriminate it from any other gesture action

can be found in Grund et al., 2023). While labour-intensive, doing so allows researchers to define gesture units flexibly, and at different levels of granularity, depending on features of interest for a particular analysis or on the characteristics of available data. While potentially powerful, there remains one substantial challenge to effectively applying this method to the analysis of gestural data: discriminating levels of lumping or splitting of gesture tokens that provide salient differences in information to the primates using them.

One hurdle in exploring gradation in gestural repertoires is that the levels of lumping and splitting vary not only between different repertoires, but also between units within them. Each signal is typically described as a set of physical characteristics—the defining movements of the gesture action plus a set of modifiers. But in some cases, modifiers are already used to split gesture actions. For example, in the literature, some units are made of gesture actions split based on duration (e.g., long and short touches; Fröhlich et al., 2016), body part (e.g., embraces with arms or feet; Liebal et al., 2004a, b), repetition (e.g., hitting once vs repeatedly; Hobaiter & Byrne, 2011a, b) or the use of one or both limbs (e.g., grabbing with one or two hands; Genty et al., 2009). At the same time, for other gesture units in the same repertoire, these same properties might be lumped and considered to be variations within the expression of a single gesture unit. Sometimes, some—but not all—units are split by their context of production (e.g., chest beat vs chest beat play, Genty et al., 2009; lunging vs lunging in play, Gupta & Sinha, 2019). In more commonly produced actions, patterns of difference that could be used to discriminate separate gesture units may be more easily detected than in rarely observed ones. But, typically, the approaches to defining a novel gesture unit via lumping and splitting are not transparent (Rodrigues et al., 2021). Importantly, there is no absolute correct level of splitting or lumping; the level of detail in a communicative repertoire depends on the question being asked. In spoken language, we construct sets of phonemes and syllables, as well as of words and higher order grammatical structures. However, it is important that a set is composed of similarly characterised units, whereas our current ape gesture repertoires are frequently composed of a mix of the equivalent of syllables and words.

Gradedness in gestures might not be as apparent to observers as it is in vocal and facial communication. However, the actions that make a gesture recognisable can be produced with different morphological features: they can be produced in combinations of different channels (acoustic, visual, tactile), with different limbs, at different intervals, at different speeds, and using different space. For example, an embrace can be defined as the signaller wrapping their limbs around the recipient, making it distinct from any other gesture action. However, whether one or both arms are used or the feet (Liebal et al., 2004a, b), which arm, which body part is embraced, for how long, etc., may still contain information about the goal of the gesture, the context of use, and/or the social relationship between signaller and recipient. While 'embraces' are clearly distinct gesture actions that most observers can reliably identify, a more fine-grained splitting process (e.g., one-armed embraces from the side around the shoulders vs bimanual embraces from behind) is possible. A recent study identified 36 different forms of the 'touch' gesture in chimpanzees, directed toward 70 different locations on the body (Bard et al., 2019). Typical modifiers include the body part with which the signaller produced the gesture; the body part on the recipient where the gesture made contact; whether the gesture was repeated; which limb was used; whether specific body parts (such as the fingers, wrist, or elbow) were flexed; and whether an object or physical substrate was involved (Bard et al., 2019; Grund et al., 2023; Hobaiter & Byrne, 2017; Roberts et al., 2012). Some coders ascertain the actions of multiple body parts for each action (e.g., Forrester, 2008), but this can be time-intensive when generating large datasets (Cartmill et al., submitted), at least until automation through pose-estimation and movement-tracking programs such as DeepLabCut become possible (Wiltshire et al., 2023). It is important to use existing large datasets to establish which level of granularity and coding effort is both useful and justifiable.

As soon as we start to compare communities or species, analyse meaning, or analyse sequential structure, the ability

to discriminate boundaries on the basis of relevant variation, to parse gesturing into discrete gestural units, becomes of substantial importance. A previous study attempted to use unsupervised cluster analysis to establish types of gestures based on morphological features of each gesture instance alone without resorting to human-defined gesture actions (Roberts et al., 2012). However, the large number of possible morphological features in relation to the small number of available data points made it impossible to interpret the results of early efforts. As gesture datasets are coded manually, small sample sizes often prevent meaningful statistical analyses, and splitting gesture actions with too much detail leads to large numbers of rare combinations of modifiers, amplifying this challenge (Bard et al., 2019). One study of chimpanzee gestures tried to define which features in a highly split repertoire were salient on the basis of whether they appeared to shift the signallers' goals (for example, whether alternate or simultaneous hitting actions led to a different outcome; Hobaiter & Byrne, 2017). A first attempt to take a species-centric perspective to defining gestural repertoires, this approach was largely descriptive. Moreover, the assessment of whether to lump up from the finest level of splitting was made based on similarity of the signaller's goal across the different units. For example, where two hit actions were made with different body parts, these would be lumped up if they achieved a similar goal or set of goals. In incorporating signaller goal as a feature that discriminated gesture 'type', did not allow us to subsequently evaluate whether the resulting signal sets were more or less 'meaningful'—in terms of goal specificity—to the chimpanzees using them (Hobaiter & Byrne, 2017).

In this paper, we address the challenge of identifying gradedness in gestural communication by designing a process that lets us split existing gesture actions at a fine scale and determine whether this splitting process increases the information content of the system. This approach builds on the bottom-up, highly detailed coding scheme employed to generate the data (Grund et al., 2023). We take a pragmatic approach: following gesture coding traditions in human co-speech and co-sign gesture (Kendon, 2004), we retain the traditional concept of a predefined set of gesture actions, as a useful shorthand summarising a large number of possible dimensions that could otherwise create a lot of noise. Similarly to the integration of handshape, location, and movement to discriminate signs in American Sign Language (Stokoe, 1960), we incorporate information from a series of modifiers. We use unsupervised clustering to split each gesture action into multiple variants based on the modifier distributions, but introduce sampling thresholds to prevent the creation of a near-infinite number of rare cases. This procedure provides us with a maximum repertoire (given the available data) that has been created without including the interaction's social context or outcome. This key information

on context and outcome can then instead be used to establish which of the variants of each gesture action, which we term 'gesture morphs' (Grund et al., 2023), represent distinct communicative signals that contain information for recipients.

Even with our understanding of the language space, our ability to classify words in new languages without input from speakers takes huge datasets (Hvitfeldt & Silge, 2022). Nevertheless, increasing computational power, larger datasets, and sophistication in machine learning and model-based analysis have offered new methods to approach this problem. Parsing words and sentence boundaries out of streams of speech data is a fundamental task of natural language processing, a task which supervised and unsupervised machine learning models are increasingly capable of solving efficiently (Chollet et al., 2022). For example, human facial movement recognition is similarly able to identify relatively discrete facial signals out of video data (Zhou et al., 2010). Applying similar approaches to animal calls has allowed for the data-driven generation of population- or species-specific repertoires, especially for highly vocal species with large training datasets available (e.g., orcas: Bergler et al., 2019; budgerigars and long-billed hermits: Keen et al., 2021; meerkats: Thomas et al., 2022). More recently, the availability of standardised software for feature extraction and analysis, such as Koe (Fukuzawa et al., 2020) and Deep-Squeak (Coffey et al., 2019), makes these tasks easier than ever where large corpora are available. Further, for species with highly graded call systems, fuzzy clustering approaches might prove more insightful in quantifying the complexity of the call repertoire (Fischer et al., 2017). Automated analyses of repertoires of animal facial signals have been introduced (Dolensek et al., 2020) but are currently limited to reactions to specific 'emotional' stimuli (such as the detection of pain; Andersen et al., 2021) and are hampered by the lack of large datasets of natural facial signals.

Cluster detection of any kind has only been applied in gesture to insufficiently large datasets (e.g., $n = 128$ tokens, Roberts et al., 2012), limiting interpretation. Cluster analysis is data-intensive, and any result is conditional on amassing (a) a sufficiently large dataset (Bouveyron et al., 2019), and (b) one that is coded in a bottom-up transparently structured way, both of which are extremely time-intensive. Gestural video data, especially when collected in non-standardised 'noisy' visual-audio environments such as forests, are higher-dimensional than sound files and videos of faces (which can be treated as two-dimensional without losing much information), introducing considerable challenges for automated feature extraction. Here we take advantage of a newly available large set of gestural data from East African chimpanzees (*Pan troglodytes schweinfurthii*): the Gestural Origins database (Hobaiter et al., 2021), generated with the GesturalOrigins coding scheme (Grund et al, 2023). We use

this dataset to test whether gestural actions, the most typical level of discrimination in a gestural repertoire, can be split further through the systematic application of modifiers, and whether such a split would add information to investigations of the function and form of gestural communication. We use latent class analysis (LCA), a statistical method used to identify unobservable subgroups based on patterns of categorical variables, to determine divisions within gesture actions that take into account the dependence structure of modifiers that occur at least five times, indicating that these splits are meaningful, replicable using other datasets, and can be used for further analysis. We hope that in doing so we offer a systematic framework for the classification of gestural data. Importantly, the level of granularity in any given study will depend on available data—the more data available (overall, and for specific gesture actions), the more splitting is possible. Using the largest existing gesture dataset, we provide a suggestion for the dimensions along which to split gesture actions, and test whether this approach can improve our understanding of gesture meaning.

Methods

Dataset

To detect the underlying structure of gesture actions, we focus on the well-studied gestural system of East African chimpanzees (*Pan troglodytes schweinfurthii*). We combined data from five long-term research communities in four sites: Sonso and Waibira communities, in the Budongo Forest Reserve, Uganda; the Issa Valley chimpanzees, in Western Tanzania; the M-group of Kalinzu Forest chimpanzees, Uganda; and the Kanyawara community, in the Kibale National Park, Uganda. Details of the observation effort and sample size for each community can be found in Table 2. Data were coded using the same protocol across communities. In total, 7879 gestures were available for which coding was complete and signaller and recipients could be identified, representing 90 distinct gesture actions before preprocessing. The dataset is biased towards the

Table 2 Description of the communities

Community	Number of gesture tokens	Number of gesture tokens—final	Number of signallers	Gesture actions
Sonso	4956	4903	75	48
Waibira	1537	1477	84	41
Issa	586	576	32	40
Kanyawara	494	493	40	30
Kalinzu	306	299	41	27

Budongo chimpanzees, which provided 82% of all gestures to the dataset, and within Budongo to chimpanzees in the Sonso community, which provide 62% of all gestures to the dataset. Gesture data were collected across a full range of behavioural contexts ($n = 22$, see Grund et al., 2023 for ethograms), but 31% of gestures occurred before or during play.

The coding scheme for this project has been described in detail in Grund et al. (2023), including definitions of all gesture actions and modifier variables employed to describe the full dataset. To study the variation in modifiers within gesture actions, we selected four of these modifiers: (i) the body part with which the gesture was produced ('body part signaller'; 11 levels); (ii) the body part with which contact was established either on the recipient or the signaller's own body ('body part contact'; nine levels); (iii) whether the gesture was repeated rhythmically or not ('repetition'; two levels); and (iv) whether the gesture was produced with one limb, both limbs at the same time, or both limbs alternatingly ('laterality'; three levels). More information about the different modifier levels and interobserver reliability tests on gesture coding can be found in the Supplementary Materials. Initially, we had included the involvement of objects in a gesture as a modifier. However, the original list of gesture actions were largely pre-split based on object use (e.g., 'hit object' was already discriminated from 'hit other' at the gesture action level); thus there was little variation within gesture actions. We had the choice of either lumping up the actions and retaining object involvement as a modifier, or splitting object involvement at the level of the gesture actions. Here we decided to do the latter, incorporating it into the gesture action level, as this tended to more closely map onto previous repertoires and simplified our set of modifiers. The relevance of object-use as a modifier might be different if this approach were applied to a different dataset, for example chimpanzees may vary their use of objects depending on their structural and acoustic properties (Fitzgerald et al., 2022; Gibson et al., 2023). Other modifiers are available in the original dataset but are often only applicable to a subset of gesture actions or cases within gesture actions (for example: finger flexion); or are rarely or never part of other primate gesture coding datasets, thereby limiting applicability of results for other users. We focused on those modifiers that are commonly coded across research groups and that affect most gesture actions. Some modifiers were originally coded with more levels than included here but were lumped based on predetermined criteria. For example, left- and right-handed unimanual gestures were combined as 'unimanual'.

In the original coding protocol, some gesture actions were pre-split based on modifiers that are of interest here, resembling predefined morphs. For example, 'hit other' indicates that the recipient received one hit, while 'hitting other' indicates that multiple hits had taken place. To reduce all gesture

actions to the same level of splitting before determining the morphs and checking whether these predefined splits were justified, we lumped gesture actions that were pre-split in the original coding based on modifiers of interest (here both would be lumped as gesture action 'hit other' with modifier repetition as yes (hitting) or no (hit)) before proceeding with the analysis, resulting in 61 gesture actions across all sites.

Preprocessing

Rare occurrences of gesture actions or modifier levels make it hard to understand the underlying usage rules—a common problem in linguistic analyses that is usually resolved by excluding rare elements (Levshina, 2015). Thus, in an initial preprocessing step, we removed 19 of the 61 gesture actions that had fewer than 10 occurrences available in the dataset, assuming that it is currently impossible to discriminate distinct morphs within them due to lack of variation and because the latent class analysis performs poorly in very small samples (Nylund-Gibson & Choi, 2018).

In addition to having a filter for gesture actions, we included several inclusion criteria for modifiers: First, within any gesture action, modifier levels that occurred fewer than five times were combined into an 'other' category (Hvitfeldt & Silge, 2022); if the 'other' category had fewer than five cases, those cases were set to missing values. Second, if a modifier was coded as 'unknown', its value was set to missing values. Third, if levels of one modifier prevent the use of another modifier (for example, gestures involving the head cannot be coded for laterality), the modifier was set to 'not valid'. Fourth, modifiers were removed within gesture actions if there was no variation—so, to include modifiers within gesture actions, we required that at least two modifier levels occurred at least five times. Individual gesture cases were removed if they had a missing value in a modifier that would otherwise have had sufficient variation to be considered for analysis within that gesture action. For a summary of modifier levels that showed sufficient cases and variation within gesture actions, see Supplementary Material.

Resulting dataset

After preprocessing, we were left with a dataset of 7752 gesture tokens across 42 gesture actions that occurred at least 10 times (range: 10–871 cases, median: 68 cases). The distribution across field sites can be found in Table 1. 'Body part signaller' and 'body part contact' were coded for all 42 gesture actions (even if the latter was regularly 'none' or 'not valid' for some gesture actions); 'laterality' was coded for 27 gesture actions; and 'repetition' was coded for 39 gesture actions.

Morphs

Gesture actions differ in the degree of modifier variation they can show. For example, given the modifiers we selected from within those available in the coded data, the gesture action 'somersault' can only be done in one specific way. For this and other simple gesture actions with one or two modifiers coded, we could just assign a morph to every combination of levels of the two modifiers. However, other gesture actions, such as 'touch' or 'hit other', have over 30 different combinations of modifier levels available. In most studies (including our own) this generates too few cases across too many categories, and it is therefore impractical for further analyses to assume that each of these distinct combinations is its own distinct signal.

In total, we had 527 distinct gesture action/modifier combinations in the dataset, which would leave us with 13 morphs of each gesture action on average. Using all possible gesture action/modifier combinations as morphs would also ignore the non-independence of modifier levels. For example, it is conceivable that repeated actions are more likely to occur when using both hands rather than one. We need an approach that can find statistically meaningful combinations of modifier levels within gesture actions that occur at least a certain number of times to rule out noise. We chose to set a threshold of five occurrences of modifier levels and morphs throughout to consider them for further analyses. Higher thresholds would reduce the number of morphs by reducing the considered number of dimensions along which clusters are detected and filtering morphs out subsequently, while a lower threshold would allow for more complex splitting choices but potentially create non-replicable morphs based on rare cases. While this gesture dataset represents the largest currently available, it remains smaller than those typically used in linguistic analysis, and thus this threshold is low as compared to standards in linguistic corpora (Levshina, 2015). The smaller datasets available in animal communication research in general and the large number of rare gesture actions and modifier levels in this dataset necessitated a generous approach at the risk of sometimes including spurious morphs.

Our goal was to identify the smallest number of gesture action/modifier combinations (morphs) to which we can accurately assign each gesture case. This last feature is important: a morph is only useful if any new case that is added to the dataset can be assigned to exactly one morph—the same way any observed gesture can be categorised as one and only one gesture action. For this purpose, we use Bayesian latent class analysis, a model-based cluster detection algorithm (Li et al., 2018), as implemented in R (R Development Core Team & R Core Team, 2020) using the 'BayesLCA' package (White & Murphy, 2016).

Latent class analysis is a statistical technique used to identify underlying subgroups from a set of observed categorical variables (Lazarsfeld & Henry, 1968). These variables are usually unordered and discrete. Our use of LCA here can be seen as the factorial equivalent to the increasing use of uniform manifold approximation and projection (UMAP) and *t*-distributed stochastic neighbour embedding (TSNE) as dimension reduction algorithms for continuous data in vocal analyses (Thomas et al., 2022). We are interested in identifying latent classes based on the modifier levels within gesture actions. The main assumption of LCA is that the observed variables are generated by a finite mixture of unobserved groups (Bouveyron et al., 2019). Each class is assumed to be mutually exclusive and exhaustive, meaning that each observation can only belong to one class. LCA is a model-based approach, which means that it uses statistical models to identify the best solution (Bouveyron et al., 2019). We assess model fit using the Bayesian information criterion (BIC), which penalises models with more parameters (Weller et al., 2020). The LCA model estimates two parameters for a given number of underlying classes: the size of each class, and the conditional probabilities of each data point to belong to each class. LCA assumes that the observed variables are conditionally independent, meaning that there is no correlation between variables once the latent class membership is known. This assumption is violated in our dataset: modifier levels can be mutually exclusive or correlated (for example, only certain body parts can be used bilaterally). Small sample sizes may lead to unstable or inaccurate class attributions in LCA (Nylund-Gibson & Choi, 2018)—one of the reasons we removed gesture actions with low number of cases. Having simple and well-separating covariates can mitigate the impact of low sample sizes (Wurpts & Geiser, 2014). Despite the violation of the conditional independence assumption and smaller than ideal sample sizes, we consider LCA a useful tool for splitting gesture actions further, but researchers should be aware that some of the found cluster solutions for the less well represented gesture actions can be unstable and may fail to replicate in new data. We assume that, with more data, more morphs would be detectable and some of the less-well represented current morphs would change, either because currently excluded modifier levels would become available or because the LCA would split clusters that are currently lumped. We also lack sufficient cases for each individual/gesture type combination to account for individual differences, potentially another source of pseudoreplication. However, the same problem (larger sample size reveals more distinct elements) applies to any repertoire analysis, and the better-represented—and thus more stable—morphs are those that also represent the majority of gestures used by chimpanzees. The number of unique modifier levels that were observed for a gesture action determine the maximum number of clusters that could be detected as they determine the number of binary dimensions along which the cluster detection takes place. Thus, a gesture action for which the coding scheme only allowed two different levels of one modifier will be limited to a maximum of two clusters, while one with variation across all modifiers could be split much more finely, given sufficient cases.

Our approach was to take all cases for a given gesture action, remove modifiers that did not show sufficient variation (see *Preprocessing*), and one-hot dummy code modifier levels so that each modifier level was represented as 0 or 1 for each gesture case (Hvitfeldt & Silge, 2022). We based the maximum number of morphs into which a gesture action could be split on the number of unique combinations of modifier levels within the gesture action. We then used LCA to determine the fit for each possible number of clusters between one and that maximum value, ten times per possible number of clusters. This last step increases robustness, as different iterations of the LCA could lead to different results, based on different starting conditions (Li et al., 2018). We extracted the BIC as a measure of model fit of the latent solution. The best cluster solution had to fulfil two conditions: a) all clusters had to be deterministic with regard to a set of modifiers, so that new cases can be unambiguously assigned to one of them; b) none of the resulting clusters included fewer than five cases (Weller et al., 2020). If multiple cluster solutions fulfilled these criteria, we chose the one with the lowest BIC.

For the chosen cluster solution, we extracted the cluster assignment for each gesture case. We subsequently tested whether all clusters within a gesture action could be explained exhaustively by any linear combination of modifier levels, by calculating the probability that each modifier level combination occurred in each cluster and the specificity of this modifier level combination to the cluster. If any modifier level combination had a probability of 1 (meaning it occurred in all cases of the cluster) and a specificity of 1 (meaning it only occurred in this cluster and no other), this modifier level combination became the defining rule for the morph. These assignments can therefore be constructed as a decision tree that gets increasingly complex. For example, the gesture action 'object shake' is split into four clusters: object shakes without repetition (with any limb); with repetition and using the foot; with repetition, using one hand; and with repetition, using both hands simultaneously. If a new case is added, it can be immediately assigned. For gesture actions with clusters that were not exhaustively described by any one modifier level combination, we checked whether several modifier level combinations reached perfect probability and specificity (for example, a cluster that combines the use of hand and fingers as body part). If this did not lead

to a reproducible ruleset to describe the morph, we left this morph as 'unexplained'. This only occurred in this data set once and was due to the LCA creating an 'all other cases' category.

Value of morphs

While it is possible to split gesture actions into morphs based on modifiers, the question remains whether a maximally split repertoire is valuable to researchers. We propose two criteria that would make a given morph valuable for current gesture research: a) if we can establish that there are clear-cut community differences in the usage of (some) morphs of the same gesture action; and/or b) if morphs of the same gesture action have different meanings and reduce uncertainty about the outcome of the interaction. We use the goal or 'apparently satisfactory outcome' of the gesture action as a proxy for meaning (following Cartmill & Byrne, 2010; Hobaiter & Byrne, 2014). All goals and their definitions can be found in the Supplementary Material. After establishing the morphs, we calculated community and goal distributions within gesture actions that have at least two morphs. We established whether a morph had less uncertainty than the gesture action in either domain by calculating the information entropy of communities and goals within each morph of a gesture action (Shannon, 1948) using the 'infotheo' package (Meyer, 2022). We then permuted the community/goal membership across morphs within a gesture action, and calculated the resulting entropies, with 1000 repetitions. This permutation procedure establishes the expected entropy of community/goal distributions if the distribution within gesture actions was random. If the observed entropy of a morph is lower than expected (i.e., smaller than at least 950 of 1000 permutations, at an alpha level of $p = 0.05$), we assume that the morph split increased our ability to correctly predict the outcome of a gesture from the morphs, rather than just the gesture action. For this analysis, we removed cases where goals could not be established, which is often the case if recipients do not openly react to a gesture or produce a reaction that is not considered a plausible goal (e.g., an attack). We make the assumption that missing goal assignment is randomly distributed across morphs within gesture actions, an assumption that was only violated in 3 out of 42 gesture actions (based on a Chi-square test for known and unknown goals within gesture actions).

As an additional way to test whether morphs increase the predictability of the community or goal, we implemented a naïve Bayes classifier using the 'naivebayes' package in R (Majka, 2019). Naïve Bayes classifiers are well established and implemented, using vectors of feature values (in our case, the gesture action or morph) to predict the correct outcome (here, the goal or community) using Bayes' theorem

(Eisenstein, 2019). The dataset was split into ten folds, the classifier was trained on 90% of the data, and the correct classification rate of the remaining 10% was noted (Hvitfeldt & Silge, 2022). We assume that if morphs are a relevant split, they should predict goals better than gesture actions alone, either across the whole dataset or within a subset of gesture actions.

Results

We included 42 gesture actions in this analysis, based on a cut-off of at least 10 cases available for each of them. Modifier levels were included into the LCA if they occurred at least five times. For 12 gesture actions, only one way of performing them was ever present in the dataset after removing rare occurrences of modifier levels (Table 3 for overview). Of the remaining 30 gesture actions, in a further five cases, the LCA extracted one cluster as the most likely solution. Thus, 17 out of 42 gesture actions were represented as a unified single morph across sites in our dataset (in addition to the 19 gesture actions for which insufficient data were available to establish morphs). The remaining gesture actions were each split into between two and seven morphs (see Table 3 for their distribution). In all, the dataset yielded 115 morphs. Table 4 contains information on all gesture actions, the number of morphs into which they were split, and the modifiers that were used to make the split. For 99 out of 115 morphs, one clear rule of modifier level combinations could be established that would allow new cases to be sorted into these categories. Of the remaining 16 morphs, 15 were combinations of rare cases; e.g., instead of splitting 'biting' based on all contact points, the rare cases of biting the face, hand, and legs were combined into one morph. After removing rare cases and gesture actions with missing information, 7560 out of 7749 gesture cases could be assigned to a single morph.

Splitting of gesture actions was primarily done based on body parts and laterality, as multiple choices were available

Table 3 Distribution of the number of morphs that the 42 gesture actions were split into

Number of morphs per gesture action	Frequency
1	17
2	6
3	7
4	4
5	1
6	5
7	2

Table 4 All morphs, and the modifier levels that define them. Dashes indicate that either of those levels would be classed in that morph. NV indicates that the modifier level is not defined within another modifier level (e.g., 'laterality' for gestures using the head). The prediction accuracy provides information about the distribution of certainty across morphs within the same gesture action

Morph name	Count	Body part Signaller	Body part contact	Repetition	Laterality	Prediction accuracy
Beckon.1	18	Arm				0
Beckon.2	12	Hand				0.25
BigLoudScratch.1	344		Body		Unimanual	0.69
BigLoudScratch.2	219		Arm		Unimanual	0.05
BigLoudScratch.3	183		Head		Unimanual	0.72
BigLoudScratch.4	74		Leg			0.06
BigLoudScratch.5	24		Back|Face		Unimanual	0.62
BigLoudScratch.6	6				Both	0
Bite.1	19		Face|Hand|Head|Leg|None			0.16
Bite.2	42		Back			0.16
Bite.3	25		Arm			0
Bite.4	25		Body			0.43
Bounce.1	14					0.42
Bow.1	12					0.17
Dangle.1	241					0.83
DangleShake.1	40					0.32
Drum.1	23					0.34
Embrace.1	34		Body		Both	0.77
Embrace.2	23		Body		Unimanual	0.65
Embrace.3	19		Back		Unimanual	0.28
Embrace.4	5		Back		Both	1
Fling.1	70	Arm				0.81
Fling.2	28	Hand				0.63
Grab.1	47		Leg		Unimanual	0
Grab.2	43		Face|Hand|Head		Unimanual	0.13
Grab.3	41		Back		Unimanual	0
Grab.4	28		Arm		Unimanual	0.08
Grab.5	26		Body		Unimanual	0.02
Grab.6	17				Both	0.25
GrabHold.1	53		Arm|Back|Body|Face|Hand|Head		Unimanual	0
GrabHold.2	31		Leg		Unimanual	0.02
GrabHold.3	25				Both	0.05
HeadStand.1	38					0.19
HitObject.1	294	Hand		No	Unimanual	0.25
HitObject.2	60	Hand		No	Both	0.13
HitObject.3	39	Foot		No		0.03
HitObject.4	34	All other cases				0
HitObject.5	32	Hand		Yes	Unimanual	0.08
HitObject.6	16	Foot		yes		0
HitOther.1	113	Hand	Arm|Body|Face|Hand|Genitals|Head|Leg	No	Unimanual	0.13
HitOther.2	55			Yes	Unimanual	0.02
HitOther.3	55	Hand	Back	No	Alternating|unimanual	0.17
HitOther.4	34			Yes	Alternating|Both	0.96
HitOther.5	30	Fingers|Foot		No	Unimanual	0.20
HitOther.6	29			No	Both	0
Jump.1	38					0

Table 4 (continued)

Morph name	Count	Body part Signaller	Body part contact	Repetition	Laterality	Prediction accuracy
KickPunch.1	10					0.2
LeafClip.1	31	Hand			Unimanual	0.77
LeafClip.2	19	Face			NV	0.95
LocomoteGallop.1	49					0.89
Lunge.1	13					0.2
ObjectMouth.1	31		s			0.14
ObjectMove.1	211				Unimanual	0.10
ObjectMove.2	37				Both	0.07
ObjectShake.1	402	Hand		Yes	Unimanual	0.51
ObjectShake.2	135	Hand		Yes	Alternating|Both	0.13
ObjectShake.3	18	Foot		Yes		0.25
ObjectShake.4	17			No		0.8
Poke.1	12					0
Present.1	197	Back				0.32
Present.2	144	Arm				0.25
Present.3	120	Body			NV	0.31
Present.4	94	Leg				0.43
Present.5	50	Bottom				0.26
Present.6	21	Head				0.75
Present.7	13	Body|Foot			Unimanual	0.1
PresentGenitals.1	283	Genitals				0.75
PresentGenitals.2	60	Bottom				0.43
Pull.1	59		Back|Body|Face|Hand		Unimanual	0.06
Pull.2	54		Leg		Unimanual	0.02
Pull.3	44		Arm		Unimanual	0.24
Pull.4	25				Both	0.24
Push.1	99	Hand	Back		Unimanual	0.19
Push.2	92	Hand	Body|Face|Hand|Head		Unimanual	0.49
Push.3	65	Hand	Leg		Unimanual	0.86
Push.4	54	Fingers				0.58
Push.5	47	Hand	Arm		Unimanual	0.15
Push.6	18				Both	0.53
Push.7	16	Foot				0.1
Raise.1	201	Arm			Unimanual	0.32
Raise.2	24	Hand				0
Raise.3	8				Both	0.38
Reach.1	370	Arm				0.37
Reach.2	191	Hand				0.70
Reach.3	16	Leg				0
Rocking.1	36					0.70
RollOver.1	34					0
Rub.1	21			No		0.43
Rub.2	19	Genitals		Yes		0.83
Rub.3	7	Bottom		Yes		0.67
Shake.1	36				Unimanual	0
Shake.2	23	Head			NV	0.55
Shake.3	16				Alternating|Both	0.87
SpinRoulade.1	23					0.47

Table 4 (continued)

Morph name	Count	Body part Signaller	Body part contact	Repetition	Laterality	Prediction accuracy
SpinSomersault.1	42					0.39
StompObject.1	213	Foot		No	Unimanual	0.06
StompObject.2	93			Yes	Alternating	0.71
StompObject.3	49			Yes	Unimanual	0.19
StompObject.4	38	Foot			Both	0.87
StompObject.5	13			No	Alternating	0.17
StompObject.6	8	Hand				0.25
StompOther.1	12		Back	No		0
StompOther.2	8			Yes		1
StompOther.3	6		Other	No		0.33
Stroke.1	16	Hand		No		0
Stroke.2	12			Yes		0
Stroke.3	8	Fingers		No		0
Swing.1	259	Arm				0.13
Swing.2	18	Leg				0.18
ThrowObject.1	20					0.33
Touch.1	144	Hand	Arm|Face|Hand|Genitals|Head			0.13
Touch.2	119	Fingers|Other				0.01
Touch.3	88	Hand	Back			0.07
Touch.4	54	Hand	Leg			0.14
Touch.5	42	Hand	Body			0.1

in many gesture actions (Fig. 1). Repetition was less relevant, because most gesture actions occurred either exclusively with repetition or without. Of the 42 gesture actions, 17 (41%) were not split based on any modifier (because there was only a single morph to the gesture action), 9 gesture actions (21%) were categorised as morphs by a single modifier, 11 gesture actions (26%) were split along two modifiers, 4 gesture actions (10%) were split by three modifiers, and 1 gestures action (2%) was split by four modifiers.

When looking at the uncertainty of the goal of morphs and gesture actions, we find that in 21 out of 25 gesture actions that were split into morphs, at least one morph existed that had significantly lower entropy (less uncertainty) than would be expected given the distribution of goals in the gesture action itself. This pattern indicates that splitting the gesture action into morphs potentially increases prediction certainty about the meaning. In 16 gesture actions, there was at least one morph that seemed to improve the probability that play behaviour was observed, indicating the existence of play-specific morphs. Throughout the dataset, 6 out of 115 morphs were exclusive to play when excluding unknown goals (two 'stomp other' morphs, one 'grab' morph, one 'shake' morph, one 'objectmouth' morph and one 'dangle' morph). Similarly, we find that 23 out of 25 gesture actions that were split into morphs had at least one morph that decreased uncertainty about the community in

which the morph was observed, indicating the potential for community specific patterns in the representation of morphs in gesture repertoires. Of those, in only two gesture actions were the morphs specific to the Sonso community suggesting that, again, any variation in entropy was not driven by sampling biases.

Using a naïve Bayes classifier, morphs (correct classification = 0.34, expected correct classification at random 0.01) are slightly worse predictors of goals overall than gesture actions (correct classification = 0.37) on the individual gesture level. However, this is not the case across goals (Fig. 2). For 16 out of 25 goals, the morphs actually provide better prediction accuracy, and on the goal level (treating the average accuracy of each goal as a data point), morphs lead to 32% correct predictions while gesture actions lead to 26% correct predictions. The difference arises because Play, which provides around one third of all data points, is more accurately predicted by the gesture action. The improvement due to using morphs for predictions includes seven goals that would never be predicted correctly if the lumped gesture actions were used as a predictor, likely because they are a secondary goal for a gesture action with one dominant meaning. After splitting the gesture actions into morphs, some of those goals (e.g., travel, mother–infant communication) are correctly predicted at relatively high rates, indicating that splitting gesture actions more finely does offer important
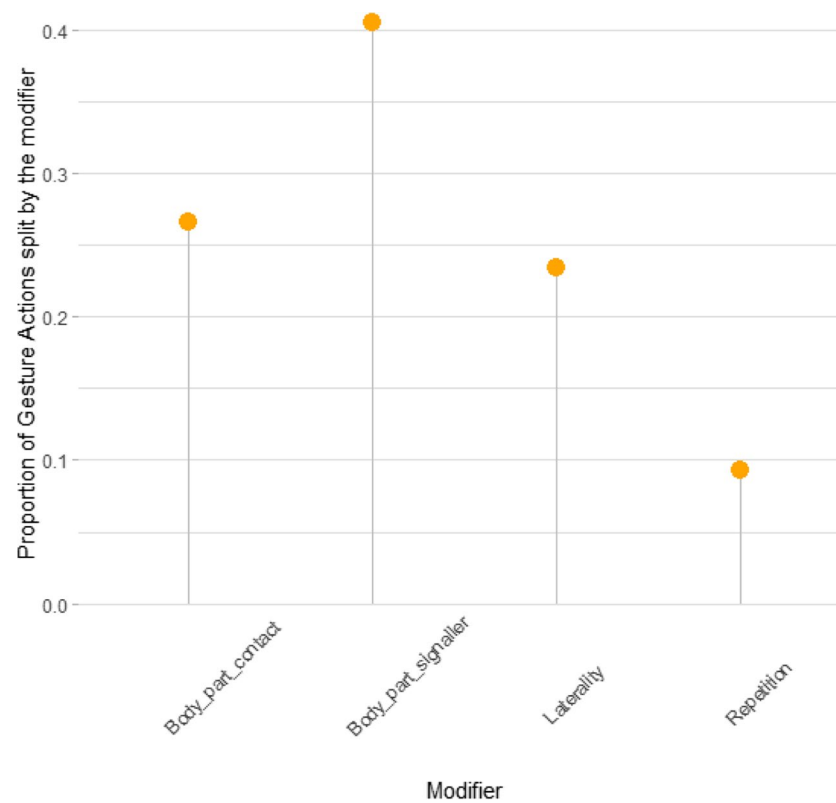
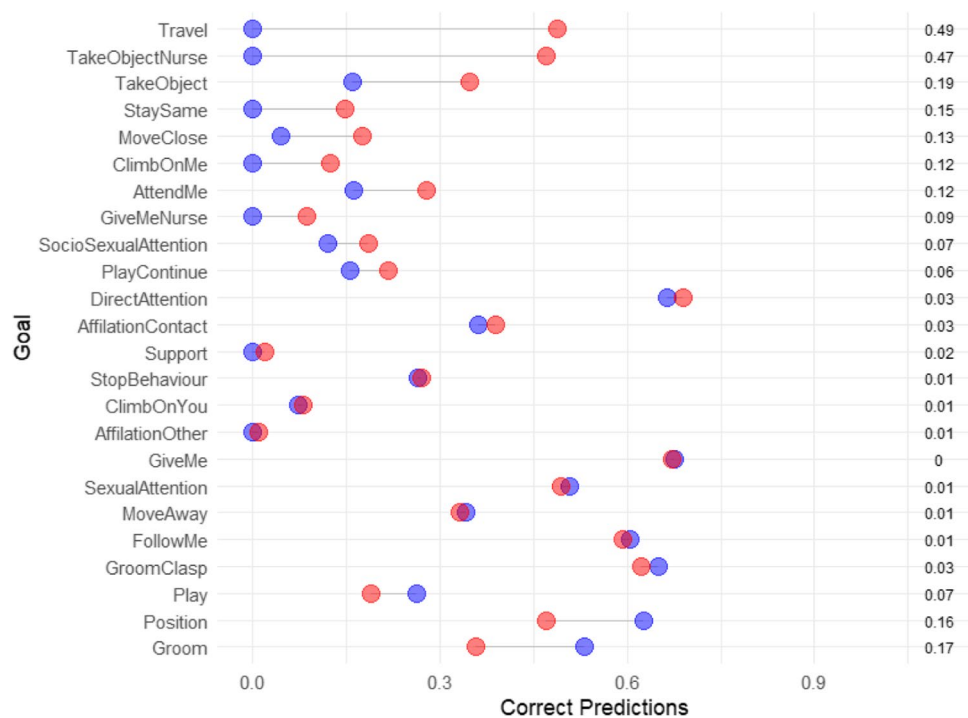**Fig. 1** Proportion of morphs that use each modifier as part of their ruleset

Fig. 2 Prediction accuracy in a naïve Bayes classifier for different goals based on gesture actions (blue) and morphs (red). Random prediction accuracy ca. 0.01.

information in those situations. For example, travel is often initiated using big loud scratch gestures (46% of travel initiations); however, the most common goal of that gesture action is to initiate grooming (66% of cases). For some of the morphs, travel is the most common goal after upsampling, so the classifier correctly assigns this target.

Discussion

One of the first tasks facing any researcher of animal communication systems is to delineate discrete elements based on their own experience, available data, and perception of the study animal. Historically, there has been a drive to start with a small number of easily discriminated elements (be they facial signals, gestures, or calls), and increasingly split them as researchers become more familiar with variation in species and individual behaviour and get better at recognising nuances that represent variation within, as compared to differences between, signal units, and as new tools become available for analysis. Through this process, signal contexts become clearer. In vocal and facial communication, recent developments in automated and manual feature extraction and classification, respectively, have allowed researchers to extend repertoires of elements or move past them to treat signals as graded and continuous on multiple dimensions (Mielke et al., 2021; Thomas et al., 2022). Here, we show that while gestural repertoires tend to have been considerably larger than those in other channels, chimpanzees may be encoding nuanced information that we are missing using typical levels of lumping in predefined gestural repertoires, justifying the process of detailed, bottom-up video coding to understand gestures as a communication system (Grund et al., 2023). We also show that the habit of lumping and splitting elements within a repertoire at different levels of granularity based on non-systematic use of modifiers potentially creates problems for comparative research. Not only does this create imbalances within a particular repertoire that may impact analysis and interpretation—much like treating an assortment of phonemes and words as a single language set—inconsistency across studies often leads to substantial sacrifices of the between-study comparison and replicability often essential to combine data from populations for the species- and family-level comparisons used for phylogenetic and evolutionary interpretation (Rodrigues et al., 2021). We argue for consistent and transparent application of lumping and splitting rules along modifying dimensions.

In this study, using an exceptionally large gesture dataset for East African chimpanzees and latent class analysis as a model-based unsupervised classifier, we show that data-based splits exist for most gesture actions that provided sufficient cases, and that the majority of the resulting 'morphs' contain information that reduced uncertainty about

the goal or chimpanzee community. Splitting here is done purely based on distributions and co-occurrences of specific modifier levels, meaning the approach is consistent across gesture actions, increasing replicability and comparability across studies and species. Thus, rather than splitting 'stomp object' based on whether it was done with one or two legs and 'hit object' based on whether it was done once or repeatedly, as was done previously in subsets of this dataset (Byrne et al., 2017; Hobaiter & Byrne, 2011a, b), both are split based on the distribution of the used body part, the repetition, and the laterality (standard modifiers in sign language coding; Kendon, 2004). However, while 'stomp object' contains alternating use of both limbs with and without repetition, 'hit object' was not observed alternatively, leading to a larger number of morphs for the former than the latter. The result is a maximally split dataset within the confines of the observed data and chosen thresholds. The number of morphs we observed (115 from 42 gesture actions with sufficient data) is considerably smaller than the number of all observed combinations of modifiers and gesture actions in the dataset (527), but all morphs occur at least five times and potentially erroneous variation in the modifiers due to coding irregularities was removed by excluding rare occurrences. If they were split, gesture actions fell into two to seven morphs, providing the basis for further investigations of usage, for example in individual or community level variation (cf., Badihi et al., 2023). The resulting set of gesture morphs allows for more variation and information than represented at the level of the gesture action repertoire, but remains small enough, with sufficient data density per unit, to handle most analyses of gestural communication.

One consistent result of ape gesture research is that—like words in human languages—gestures show means-ends dissociation (Tomasello et al., 1994; Hobaiter & Byrne, 2014): most gesture actions are used for multiple goals, and most goals can be achieved using multiple gesture actions (Graham et al., 2020). Another common result is that there is large overlap across ape species, subspecies, populations, and communities in existing gesture actions and their usage (Byrne et al., 2017; Graham et al., 2018). However, both these findings are based on the assumption that all gesture instances (often termed tokens) that are assigned to one specific gesture action are, in fact, the same communicative signal. Splitting the gesture actions into morphs could potentially reveal differences in meaning and community patterns, enabling us to get a more nuanced picture of primate gesture usage, particularly where a more systematic approach to categorisation allows us to minimise structural coding biases towards the most well-studied species (in primate gesture, typically chimpanzees; Grund et al., 2023; Rodrigues et al., 2021). Here, using morphs rather than gesture actions reduces the uncertainty about the goals and communities associated with some gesture actions: they become more

predictable. This is a strong indicator that, while gesture actions see widespread use across meanings and communities, morphs are more specific and potentially a more relevant unit of analysis for some analyses (e.g., Byrne et al., 2017). This leads to improved predictions for many goals when using a naïve Bayes classifier to predict the meaning from the gesture action or morph alone. Crucially, this is not the case across all gesture actions—for around half of them, even though morphs reduce entropy, they do not improve predictability or even add noise when using only one predictor (morph or gesture action, respectively) and a fairly simple classifier that puts high value on the most common observed target to maximise predictions. Disentangling the relationships between morphs or gesture actions and meaning or community membership will be an important task for future research, as will be the question of whether the reduced uncertainty of some morphs is driven by specific modifiers. This finding is equivalent to recent studies showing the depth of chimpanzee cultural diversity by further refining coding schemes for tool-use behaviours (Boesch et al., 2020). It remains a key step forward to recognise that, for many research questions, gesture actions might be insufficiently split to make conclusive statements about variation and flexibility of use.

Importantly, most gesture studies will not have the same level of detail or sample size as this one and may need to lump morphs or even gesture actions (rather than splitting) in order to reach a sufficiently large sample per gesture unit. Even in our dataset, around one third of all coded gesture actions did not occur at least 10 times, the threshold we set for the inclusion of gesture actions in the analysis. What then, can we achieve by splitting the repertoire ever more finely? By showing that modifiers matter and gesture actions can be split more finely, we hope to encourage researchers to code their original data with additional detail and increase the use of replicable and comparable categories and ethograms, which will make the combination and re-analysis of datasets easier (Cartmill & Hobaiter, 2019; Grund et al., 2023; Rodrigues et al., 2021). Further, it is crucial to reiterate that there is no 'correct' level of granularity in a gesture repertoire. In every communication dataset, there will be some elements that are common enough to split, potentially allowing more nuanced analyses for that subset of data. There will be some research questions that are driven by variation at higher levels of lumping—for example, looking at the use of different channels of information (tactile, visual, auditory) does not require splitting, even by gesture action (e.g., Liebal et al., 2004a, b; Gupta & Sinha, 2019; Dafreville et al., 2021). As long as research methods remain transparent about the level of analysis taking place, fine-grained coded data can be reconstituted into different categories driven by the questions of interest. To an extent, this post-hoc re-classification at a more lumped level already

occurs due to data restrictions, with researchers lumping or excluding rare elements where analyses would otherwise become untenable. In the case of dataset driven decisions, we advocate for an approach that creates the 'most-split' dataset possible (in the current case, our morphs), and, where lumping is needed, this should occur following clearly defined rules that are established a priori to coding.

An important consideration of our current repertoire of morphs is that they remain—like all repertoires—based on a set of decisions we made at the level of the coding scheme (for example, the types and splitting of modifiers that we included) and at the level of the current analysis (for example, the modifiers included and chosen thresholds). Thus, while substantially *more* bottom-up than other approaches, even more fine-grained levels of splitting are theoretically possible, as would be the inclusion of different modifiers (given a sufficient dataset). Because the latent class analysis is model-based and dependent on initial conditions, where modifier levels are rare, individual morph assignments may change between runs. In practice, however, more fine-grained coding is extremely unlikely across whole gesture repertoires because of the labour-intensive nature of coding and the size of the datasets required. Even with promising new methods to automate the detection and description of ape body-posture within video (e.g., Wiltshire et al., 2023), full automation of gesture coding in video of wild apes remains a substantial challenge. Perhaps more importantly, the continued refinement of gestural repertoires as we continue to study ape gestures and learn more is the scientific process functioning as it should. As our understanding of ape gestures becomes more refined, and as more data become available, it is likely that we will detect more morphs that currently fall below the threshold. This is akin to increasing sample sizes in any description of animal behaviour—but the crucial difference here is that the splitting process can be automatised and repeated following an established set of rules.

Few single studies of animal gestural communication will have sample sizes that allow researchers to split their gesture repertoire post hoc across all elements; however, research groups with combined corpuses or researchers focusing on common signals might be able to replicate this approach. Doing so would be valuable even if modifying variables differ from the ones used here, because it allows us to further map the variability in animal communication, similar to studies of fuzziness in vocal communication (Fischer et al., 2017). We have made all R scripts that were used to conduct analyses available in the associated GitHub repository to encourage replication of this approach. It is hard to determine the minimum sample size necessary for this effort, because it would be based on gesture action frequencies and the number of modifiers under consideration. There is no correct level of splitting or repertoire granularity; just as phonemes, syllables, and

words represent valid levels of linguistic analysis, gesture actions, gesture morphs, and even more finely split units are all valid levels of gesture analysis. But, just as in the study of languages, it is critical (a) that repertoires are composed of units that are split consistently and by features that are salient (based on recipient responses) to their *users*—here, East African chimpanzees, and (b) that researchers tailor their use of different repertoires to their question. We implemented a number of thresholds (only investigate gesture actions that occur more than 10 times; only consider morphs that occur at least five times) in an attempt to create meaningful, reproducible splits while still exploring potentially rare morphs. For many studies, having much larger thresholds would potentially be more appropriate—splitting only signals that occur at least 100 times and splitting them into a small number of morphs to avoid the creation of rare elements that then introduce uncertainty for subsequent analyses.

Important claims about the semantic flexibility of great ape gesture and its similarity to human word use (Tomasello et al., 1994; Hobaiter & Byrne, 2014) are based on the assumption that the repertoires applied represent a level of splitting suitable for unit-meaning mapping. Conversely, there remains no evidence for syntactic structures or combinatoriality in ape gesture, despite substantial research effort (Liebal et al., 2004b; Genty & Byrne, 2010; Hobaiter & Byrne, 2011b; Graham et al., 2020; however, see Oña et al., 2019, for multimodal combinations), but the detection of structural rules is only possible where signals are parsed into relevant units. In human language, variation in the tone of word production may reflect variation in emphasis or affect (as in English or Arabic) or may fundamentally change word meaning (as in Thai or Cantonese). In gesture, modifiers such as rhythmic repetition, the body part involved in production, or the inclusion and modification of objects may be used to vary emphasis or to change meaning (Kendon, 2004). The systematic description and investigation of modifiers of ape gesture production is an essential step in our understanding of gesture as a communication system. Here, we build on a newly described approach to gesture coding (Grund et al., 2023) that allows us to apply novel analytical approaches to the construction of gesture units (at the level of gesture actions and gesture morphs) derived from the apes' own gesture usage. We use the largest dataset of East African chimpanzee gesture currently available to define gesture units and repertoires that can be applied to other, smaller, datasets. We show that previously undescribed levels of granularity (gesture morphs) appear to provide additional specificity when exploring gesture meaning and may allow nuanced description of community-level variation in gesture use. In doing so we provide a foundation to develop the study of ape gesture and to redefine comparisons with other communication systems such as human language.

Data availability All scripts and data can be found here: https://github.com/AlexMielke1988/Morph_Repertoire. Data have been partially anonymised to allow for computational replication; for non-anonymised data, please contact the first or last author.

References

Andersen, P. H., Broomé, S., Rashid, M., Lundblad, J., Ask, K., Li, Z., ..., Kjellström, H. (2021). Towards machine recognition of facial expressions of pain in horses. *Animals*, *11*(6), Article 6. https://doi.org/10.3390/ani11061643

Austin, G. (1806). *Chironomia, or a treatise on rhetorical delivery*. Bulmer.

Badihi, G., Graham, K. E., Fallon, B., Safryghin, A., Soldati, A., Zuberbühler, K., & Hobaiter, C. (2023). Dialects in leaf-clipping and other leaf-modifying gestures between neighbouring communities of East African chimpanzees. *Scientific Reports, 13*(1), 147. https://doi.org/10.1038/s41598-022-25814-x

Bard, K. A., Maguire-Herring, V., Tomonaga, M., & Matsuzawa, T. (2019). The gesture 'Touch': Does meaning-making develop in chimpanzees' use of a very flexible gesture? *Animal Cognition, 22*(4), 535–550. https://doi.org/10.1007/S10071-017-1136-0/TABLES/9

Bateson, M., & Martin, P. (2021). *Measuring Behaviour*. Cambridge University Press.

Bergler, C., Schröter, H., Cheng, R. X., Barth, V., Weber, M., Nöth, E., ..., Maier, A. (2019). ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning. *Scientific Reports*, 9(1), Article 1. https://doi.org/10.1038/s41598-019-47335-w

Boesch, C., Kalan, A. K., Mundry, R., Arandjelovic, M., Pika, S., Dieguez, P., ... & Kühl, H. S. (2020). Chimpanzee ethnography reveals unexpected cultural diversity. Nature Human Behaviour, 4(9), 910-916.

Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-based clustering and classification for data science: With Applications in R*. Cambridge University Press.

Byrne, R. W., Cartmill, E., Genty, E., Graham, K. E., Hobaiter, C., & Tanner, J. (2017). Great ape gestures: Intentional communication with a rich set of innate signals. *Animal Cognition, 20*(4), 755–769. https://doi.org/10.1007/s10071-017-1096-4

Cartmill, E., & Byrne, R. W. (2010). Semantics of primate gestures: Intentional meaning of orangutan gestures. *Animal Cognition, 13*(6), 793–804.

Cartmill, E. A., & Hobaiter, C. (2019). Gesturing towards the future: Cognition, big data, and the future of comparative gesture research. *Animal Cognition, 22*(4), 597–604. https://doi.org/10.1007/s10071-019-01278-x

Chollet, F., Kalinowski, T., & Allaire, J. J. (2022). Deep Learning in R. In *R-bloggers* (2nd ed., Issue 7080). Manning Publications. https://www.manning.com/books/deep-learning-with-r

Clark, P., Waller, B. M., Burrows, A. M., Julle- Danière, E., Agil, M., Engelhardt, A., & Micheletta, J. (2020). *American Journal of Physical Anthropology, 173*(3), 411–422. https://doi.org/10.1002/ajpa.24129

Clay, Z., Archbold, J., & Zuberbühler, K. (2015). Functional flexibility in wild bonobo vocal behaviour. *PeerJ, 2015*(8), e1124. https://doi.org/10.7717/PEERJ.1124/SUPP-3

Coffey, K. R., Marx, R. E., & Neumaier, J. F. (2019). DeepSqueak: A deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology, 44*(5), Article 5. https://doi.org/10.1038/s41386-018-0303-6

Crockford, C., Gruber, T., & Zuberbühler, K. (2018). Chimpanzee quiet hoo variants differ according to context. *Royal Society Open Science, 5*(5). https://doi.org/10.1098/RSOS.172066

Dafreville, M., Hobaiter, C., Guidetti, M., Sillam-Dussès, D., & Bourjade, M. (2021). Sensitivity to the communicative partner's attentional state: A developmental study on mother–infant dyads in wild chimpanzees (Pan troglodytes schweinfurthii). *American Journal of Primatology, 83*(12), e23339. https://doi.org/10.1002/ajp.v83.12

Dolensek, N., Gehrlach, D. A., Klein, A. S., & Gogolla, N. (2020). Facial expressions of emotion states and their neuronal correlates in mice. *Science, 368*(6486), 89–94. https://doi.org/10.1126/science.aaz9468

Eisenstein, J. (2019). *Introduction to Natural Language Processing*. MIT Press.

Fitzgerald, M., Willems, E. P., Gaspard Soumah, A., Matsuzawa, T., & Koops, K. (2022). To drum or not to drum: Selectivity in tree buttress drumming by chimpanzees (*Pan troglodytes verus*) in the Nimba Mountains, Guinea. *American Journal of Primatology*. e23382. https://doi.org/10.1002/ajp.23382

Fischer, J., Wadewitz, P., & Hammerschmidt, K. (2017). Structural variability and communicative complexity in acoustic communication. *Animal Behaviour, 134*, 229–237. https://doi.org/10.1016/j.anbehav.2016.06.012

Forrester, G. S. (2008). A multidimensional approach to investigations of behaviour: Revealing structure in animal communication signals. *Animal Behaviour, 76*(5), 1749–1760. https://doi.org/10.1016/j.anbehav.2008.05.026

Fröhlich, M., Wittig, R. M., & Pika, S. (2016). Play-solicitation gestures in chimpanzees in the wild: Flexible adjustment to social circumstances and individual matrices. *Royal Society Open Science, 3*(8), e160278. https://doi.org/10.1098/rsos.160278

Fukuzawa, Y., Webb, W. H., Pawley, M. D. M., Roper, M. M., Marsland, S., Brunton, D. H., & Gilman, A. (2020). Koe: Web-based software to classify acoustic units and analyse sequence structure in animal vocalizations. *Methods in Ecology and Evolution, 11*(3), 431–441. https://doi.org/10.1111/2041-210X.13336

Genty, E., Breuer, T., Hobaiter, C., & Byrne, R. W. (2009). Gestural communication of the gorilla (Gorilla gorilla): Repertoire, intentionality and possible origins. *Animal Cognition, 12*(3), 527–546. https://doi.org/10.1007/s10071-009-0213-4

Genty, E., & Byrne, R. W. (2010). Why do gorillas make sequences of gestures? *Animal Cognition, 13*(2), 287–301.

Gibson, V., Boysen, S. T., Hobaiter, C., & Davila-Ross, M. (2023). Object use in communication of semi-wild chimpanzees. *Animal Cognition*.

Graham, K. E., Furuichi, T., & Byrne, R. W. (2020). Context, not sequence order, affects the meaning of bonobo (Pan paniscus) gestures. *Gesture, 19*(2–3), 335–364. https://doi.org/10.1075/GEST.19028.GRA/CITE/REFWORKS

Graham, K. E., Hobaiter, C., Ounsley, J., Furuichi, T., & Byrne, R. W. (2018). Bonobo and chimpanzee gestures overlap extensively in meaning. *PLoS Biology, 16*(2). https://doi.org/10.1371/journal.pbio.2004825

Grund, C., Badihi, G., Graham, K. E., Safryghin, A., & Hobaiter, C. (2023). GesturalOrigins: A bottom-up framework for establishing systematic gesture data across ape species. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02082-9

Gupta, S., & Sinha, A. (2019). Gestural communication of wild bonnet macaques in the Bandipur National Park. *Southern India. Behavioural Processes, 168*, 103956.

Hey, J., Waples, R. S., Arnold, M. L., Butlin, R. K., & Harrison, R. G. (2003). Understanding and confronting species uncertainty in biology and conservation. *Trends in Ecology & Evolution, 18*(11), 597–603. https://doi.org/10.1016/j.tree.2003.08.014

Hobaiter, C., & Byrne, R. W. (2011a). The gestural repertoire of the wild chimpanzee. *Animal Cognition, 14*(5), 745–767. https://doi.org/10.1007/s10071-011-0409-2

Hobaiter, C., & Byrne, R. W. (2011b). Serial gesturing by wild chimpanzees: its nature and function for communication. *Animal Cognition, 14*, 827–838.

Hobaiter, C., & Byrne, R. W. (2014). The meanings of chimpanzee gestures. *Current Biology, 24*(14), 1596–1600. https://doi.org/10.1016/j.cub.2014.05.066

Hobaiter, C., & Byrne, R. W. (2017). What is a gesture? A meaning-based approach to defining gestural repertoires. *Neuroscience and Biobehavioral Reviews, 82*, 3–12. https://doi.org/10.1016/j.neubiorev.2017.03.008

Hobaiter, C., Badihi, G., Bezerra de Melo Daly, G., Eleuteri, V., Graham, K. E., Grund, C., ..., Wiltshire, C. (2021). The Great Ape Dictionary video database (1.0.0). *Zenodo*.

Hvitfeldt, E., & Silge, J. (2022). *Supervised Machine Learning for Text Analysis in R* (1st ed.). CRC Press.

Jack, R. E., Garrod, O. G. B., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology, 24*(2), 187–192. https://doi.org/10.1016/j.cub.2013.11.064

Keen, S. C., Odom, K. J., Webster, M. S., Kohn, G. M., Wright, T. F., & Araya-Salas, M. (2021). A machine learning approach for classifying and quantifying acoustic diversity. *Methods in Ecology and Evolution, 12*(7), 1213–1225. https://doi.org/10.1111/2041-210X.13599

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In the relationship of verbal and nonverbal communication. No 25. Walter de Gruyter. p.207-227

Kendon, A. (2004). Gesture: Visible Action as Utterance. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511807572

Ladygina-Kohts, N. N. & Vekker, B. T. (2002). *Infant chimpanzee and human child: A classic 1935 comparative study of ape emotions and intelligence*. Ed: de Waal, F. B. M. Oxford University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin.

Levshina, N. (2015). *How To Do Linguistics with R*. John Benjamins Publishing Company.

Li, Y., Lord-Bessen, J., Shiyko, M., & Loeb, R. (2018). Bayesian Latent Class Analysis Tutorial. *Multivariate Behavioral Research, 53*(3), 430–451. https://doi.org/10.1080/00273171.2018.1428892

Liebal, K., Pika, S., & Tomasello, M. (2004a). Social communication in siamangs (Symphalangus syndactylus): Use of gestures and facial expressions. *Primates, 45*(1), 41–57. https://doi.org/10.1007/s10329-003-0063-7

Liebal, K., Call, J., & Tomasello, M. (2004b). Use of gesture sequences in chimpanzees. *American Journal of Primatology, 64*(4), 377–396.

Majka, M. (2019). *naivebayes: High Performance Implementation of the Naive Bayes Algorithm in R. R package version 0.9.*

Marler, P. (1976). Social organization, communication and graded signals: The chimpanzee and the gorilla. In *Growing Points in Ethology*. Cambridge University Press.

McNeill, D. (1992). Hand and mind: What gestures reveal about thought. University of Chicago Press.

Meyer, P. E. (2022). *infotheo: Information-Theoretic Measures* (1.2.0.1). https://CRAN.R-project.org/package=infotheo

Mielke, A., Waller, B. M., Pérez, C., Rincon, A. V., Duboscq, J., & Micheletta, J. (2021). NetFACS: Using network science to understand facial communication systems. *Behavior Research Methods, 2021*, 1–16. https://doi.org/10.3758/S13428-021-01692-5

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science, 4*(4), 440.

Oña, L. S., Sandler, W., & Liebal, K. (2019). A stepping stone to compositionality in chimpanzee communication. *PeerJ, 7*, e7623.

Parr, L. A., Cohen, M., & de Waal, F. B. M. (2005). Influence of social context on the use of blended and graded facial displays in chimpanzees. *International Journal of Primatology, 26*(1), 73–103. https://doi.org/10.1007/s10764-005-0724-z

Parr, L. A., & Waller, B. M. (2006). Understanding chimpanzee facial expression: Insights into the evolution of communication. *Social Cognitive and Affective Neuroscience, 1*(3), 221–228. https://doi.org/10.1093/scan/nsl031

Plooij, F. X. (1979). Some Basic Traits of Language in Wild Chimpanzees. In *Action, Gesture and Symbol The Emergence of Language,* (pp. 111–131). Academic Press. https://www.scirp.org/(S(lz5mqp453edsnp55rrgjct55.))/reference/referencespapers.aspx?referenceid=2336207

R Development Core Team & R Core Team. (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing Vienna Austria, 0*, {ISBN} 3-900051-07-0. https://doi.org/10.1038/sj.hdy.6800737

Roberts, A. I., Vick, S.-J., Roberts, S. G. B., Buchanan-Smith, H. M., & Zuberbühler, K. (2012). A structure-based repertoire of manual gestures in wild chimpanzees: Statistical analyses of a graded communication system. *Evolution and Human Behavior, 33*(5), 578–589. https://doi.org/10.1016/j.evolhumbehav.2012.05.006

Rodrigues, E. D., Santos, A. J., Veppo, F., Pereira, J., & Hobaiter, C. (2021). Connecting primate gesture to the evolutionary roots of language: A systematic review. *American Journal of Primatology, 83*(9), e23313. https://doi.org/10.1002/ajp.23313

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27–52. https://doi.org/10.1016/S0010-0277(98)00075-4

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*(4), 606–621. https://doi.org/10.1006/jmla.1996.0032

Sainburg, T., Thielk, M., & Gentner, T. Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLOS Computational Biology, 16*(10), e1008228. https://doi.org/10.1371/JOURNAL.PCBI.1008228

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Slocombe, K. E., Waller, B. M., & Liebal, K. (2011). The language void: The need for multimodality in primate communication research. *Animal Behaviour, 81*(5), 919–924. https://doi.org/10.1016/j.anbehav.2011.02.002

Slocombe, K. E., & Zuberbühler, K. (2005). Agonistic screams in wild chimpanzees (Pan troglodytes schweinfurthii) vary as a function of social role. *Journal of Comparative Psychology, 119*(1), 67.

Slocombe, K. E., & Zuberbühler, K. (2007). Chimpanzees modify recruitment screams as a function of audience composition. *Proceedings of the National Academy of Sciences, 104*(43), 17228–17233.

Slocombe, K. E., Townsend, S. W., & Zuberbühler, K. (2009). Wild chimpanzees (Pan troglodytes schweinfurthii) distinguish between different scream types: evidence from a playback study. *Animal cognition, 12*, 441–449.

Slocombe, K. E., Kaller, T., Call, J., & Zuberbühler, K. (2010). Chimpanzees extract social information from agonistic screams. *PLoS One, 5*(7), e11473.

Stokoe, W. C. (1960). Sign language structure: An outline of the visual communication systems of the American deaf. In *Studies in Linguistics, Occasional Papers* (Vol. 8). University of Buffalo.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50*(1), 86–132. https://doi.org/10.1016/j.cogpsych.2004.06.001

Thomas, M., Jensen, F. H., Averly, B., Demartsev, V., Manser, M. B., Sainburg, T., ..., Strandburg-Peshkin, A. (2022). A practical guide for generating unsupervised, spectrogram-based latent space representations of animal vocalizations. *Journal of Animal Ecology, 91*(8), 1567–1581. https://doi.org/10.1111/1365-2656.13754

Tomasello, M., George, B. L., Kruger, A. C., Jeffrey, M., & Farrar, & Evans, A. (1985). The development of gestural communication in young chimpanzees. *Journal of Human Evolution, 14*(2), 175–186. https://doi.org/10.1016/S0047-2484(85)80005-1

Tomasello, M., Call, J., Nagell, K., Olguin, R., & Carpenter, M. (1994). The learning and use of gestural signals by young chimpanzees – a trans-generational study. *Primates, 35*, 137–154.

Wadewitz, P., Hammerschmidt, K., Battaglia, D., Witt, A., Wolf, F., & Fischer, J. (2015). Characterizing Vocal Repertoires—Hard vs. *Soft Classification Approaches. PLoS ONE, 10*(4), e0125785. https://doi.org/10.1371/journal.pone.0125785

Waller, B. M., & Micheletta, J. (2013). Facial expression in nonhuman animals. *Emotion Review, 5*(1), 54–59. https://doi.org/10.1177/1754073912451503

Waller, B. M., Whitehouse, J., & Micheletta, J. (2017). Rethinking primate facial expression: A predictive framework. *Neuroscience and Biobehavioral Reviews, 82*, 13–21. https://doi.org/10.1016/j.neubiorev.2016.09.005

Weller, B. E., Bowen, N. K., & Faubert, S. J. (2020). Latent Class Analysis: A Guide to Best Practice. *Journal of Black Psychology, 46*(4), 287–311. https://doi.org/10.1177/0095798420930932

White, A., & Murphy, T. B. (2016). BayesLCA : An R Package for Bayesian Latent Class Analysis. *Journal of Statiscal Software, 61*(13), 1–28. https://doi.org/10.18637/JSS.V061.I13

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology, 7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wiltshire, C., Lewis-Cheetham, J., Komedová, V., Matsuzawa, T., Graham, K. E., & Hobaiter, C. (2023). DeepWild: application of the pose estimation tool DeepLabCut for behaviour tracking in wild chimpanzees and bonobos. *Journal of Animal Ecology*. https://doi.org/10.1111/1365-2656.13932

Wurpts, I. C., & Geiser, C. (2014). Is adding more indicators to a latent class analysis beneficial or detrimental? Results of a Monte-Carlo study. *Frontiers in Psychology, 5*, 920. https://doi.org/10.3389/fpsyg.2014.00920

Zhou, F., De la Torre, F., & Cohn, J. F. (2010). Unsupervised discovery of facial events. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010*, 2574–2581. https://doi.org/10.1109/CVPR.2010.5539966