DISCOURSE PROCESSES
2022, VOL. 59, NO. 3, 209–236
https://doi.org/10.1080/0163853X.2021.1992235

Routledge
Taylor & Francis Group

The Primacy of Multimodal Alignment in Converging on Shared Symbols for Novel Referents

Marlou Rasenberg [a,b,e], Asli Özyürek [a,b,c,e], Sara Bögels [c,d,e], and Mark Dingemanse [a,b,e]

[a]Centre for Language Studies, Radboud University; [b]Max Planck Institute for Psycholinguistics; [c]Donders Institute for Brain, Cognition and Behaviour, Radboud University; [d]Department of Communication and Cognition, Tilburg University; [e]Communicative Alignment in Brain and Behaviour team, Language in Interaction consortium, the Netherlands

ABSTRACT
When people interact to establish shared symbols for novel objects or concepts, they often rely on multiple communicative modalities as well as on alignment (i.e., cross-participant repetition of communicative behavior). Yet these interactional resources have rarely been studied together, so little is known about if and how people combine multiple modalities in alignment to achieve joint reference. To investigate this, we systematically track the emergence of lexical and gestural alignment in a referential communication task with novel objects. Quantitative analyses reveal that people frequently use a combination of lexical and gestural alignment, and that such multimodal alignment tends to emerge earlier compared to unimodal alignment. Qualitative analyses of the interactional contexts in which alignment emerges reveal how people flexibly deploy lexical and gestural alignment in line with modality affordances and communicative needs.

Introduction

Even when sharing a common language, we sometimes talk about things for which we do not have conventional labels, such as abstract ideas, new innovations, or unfamiliar objects. How do people create shared symbols to refer to these novel referents? Here, we study this question in the context of multimodal interaction, the natural ecology of human language. Our aim is to understand when and how people converge on referential expressions and how they use spoken and gestural resources in this process. We focus on the interplay between two key interactional processes that are known to underlie the emergence of novel symbols: alignment (i.e., cross-participant repetition of communicative behavior) and the flexible deployment of communicative affordances of the vocal (e.g., speech) and manual (e.g., gesture) modalities.

The importance of alignment for collaborative referring to (novel) objects or concepts has been substantiated in work on alignment. People have been shown to perform better in joint cooperative tasks (such as the Map Task; Brown et al., 1984) when they align their communicative behaviors, such as lexical and syntactic choice (Dideriksen et al., 2020; Fusaroli & Tylén, 2016; Reitter & Moore, 2014). There is also evidence for a causal effect of alignment on the process of creating shared symbols: in a study involving drawings, communicative success (that is, how accurately matchers were able to identify the correct meaning based on a drawing) was higher when participants were allowed to make their drawings alike, compared to when they were forbidden to do so (Fay et al., 2018, Experiment 2).

CONTACT Marlou Rasenberg marlou.rasenberg@mpi.nl Max Planck Institute for Psycholinguistics, P.O. Box 310, Nijmegen 6500 AH, The Netherlands

The different affordances of the vocal and manual modalities for symbol creation have been a key topic in the field of language evolution or emergence (e.g., Goldin-Meadow, 2017; Levinson & Holler, 2014). When people cannot rely on conventionalized symbols to refer to (novel) objects or concepts, gestures are effective because of their iconic potential (Fay et al., 2013, 2014; Macuch Silva et al., 2020; Zlatev et al., 2017), and may therefore help "bootstrap" a communication system (Fay et al., 2013). However, when people are faced with unfamiliar stimuli, there is also evidence for a multimodal advantage of combining gestures and non-linguistic vocalizations (i.e., non-word sounds) compared to using either of those modalities alone (Macuch Silva et al., 2020), which implies that their joint contribution might facilitate shared symbol creation.

So, previous work has revealed that behavioral alignment plays a key role in collaborative referring, and that the manual modality (in combination with the vocal modality) can be used effectively for establishing joint reference to (novel) objects or concepts. Yet, we know very little about how people use the communicative affordances of multiple modalities in the process of alignment in emergence contexts. This is because alignment has mostly been studied in terms of just lexical choice or co-speech gesture, without looking at the relation between modalities, and because studies of language emergence have rarely focused on the analysis of cross-modal alignment in interactive contexts. There is a missing link in our understanding of the interplay between alignment and the affordances of communicative modalities: how do people deploy alignment in one or multiple modalities when referring to novel referents?

Here, we aim to provide a first step toward answering this question by looking at lexical and gestural alignment in a multimodal corpus of dyads performing a referential communication task with novel objects (similar to the Tangram task (Clark & Wilkes-Gibbs, 1986) but in a face-to-face setting, see Figures 1 and 2). Our primary focus is on the *emergence* of alignment: we examine the first time speakers repeat each other's lexical choice and/or gesture (i.e., align) when referring to a particular referent in a conversational context. We quantify *how often* and *when* this happens and in which modality or modalities (i.e., lexically, gesturally, or in both modalities). If alignment is established in both modalities for a particular referent (i.e., multimodal alignment), we ask next whether it emerged simultaneously (lexical and gestural alignment emerge at the same time) or successively (alignment in one modality preceding alignment in the other modality). To investigate *how* alignment is employed for collaborative referring, we qualitatively inspect its turn-by-turn unfolding and the affordances of the spoken and gestural modalities as they are recruited by participants.

Modality and alignment

A key element of the process of achieving collaborative reference is for participants to establish a shared conceptualization: a *conceptual pact*. Such conceptual pacts can be encoded in particular verbal expressions (Brennan & Clark, 1996), but also gestures (Holler & Wilkin, 2011) or drawings (Fay et al., 2018). For example, communicators can align on lexical items such as "ice skater" to refer to a Tangram figure (Clark & Wilkes-Gibbs, 1986), "line" to refer to a particular part of a maze (Garrod & Anderson, 1987), or "loafer" to refer to one out of multiple shoes (Brennan & Clark, 1996). When used repeatedly over time, such conceptual pacts are considered to have become "entrained" (Brennan & Clark, 1996).

Conceptual pacts do not appear out of the blue; they take interactional work: "speakers and addressees work together in the making of a definite reference" (Clark & Wilkes-Gibbs, 1986, p. 1). During this collaborative process (known as *grounding*), repetition of lexical choice can be particularly useful; it can be employed to accept a referring expression (Clark & Brennan, 1991), or to repair or expand it (Clark & Brennan, 1991; Clark & Wilkes-Gibbs, 1986; Dingemanse et al., 2015; Fusaroli et al., 2017). Co-speech gestures can be effective for this process as well (Chui, 2014; Holler & Wilkin, 2011; Tabensky, 2001). For example, in one study by Holler and Wilkin (2011), a participant playing the role of director described a Tangram figure with the verbal expression "with two things sticking out" along with a co-speech gesture where her two arms represent the position of the figure's arms sticking out from the back. The matcher replied with "yeah" while repeating the gesture, which signaled in a "definite manner that the entirety of the reference has been understood" (p. 143).

Holler and Wilkin's (2011) results show how lexical and gestural alignment can be recruited jointly or separately in various ways when forming conceptual pacts. They can go together, for example, when speakers both produce the lexical phrase "the ice skater," as well as the same gestural representation of the figure. Or they can part ways, as when a director said "an ostrich" together with a reenactment of the figure and the matcher replied with "Yeah, okay that, that looks like a woman to me, kicking her leg up behind her, yeah?," while producing the same gesture. Here, the matcher repeated the iconic gesture while replacing the verbal expression with an alternative conceptualization. Other work, too, shows examples where speakers copy each other's gestures in casual conversation, either with or without lexical alignment (Chui, 2014; de Fornel, 1992; Graziano et al., 2011; Kimbara, 2006; Bertrand et al., 2013).

If we were to make predictions about how frequently lexical and gestural alignment co-occur, we could expect prevalence of multimodal alignment based on the interactive alignment model by Pickering and Garrod (2004). Here, alignment is considered to be the result of linguistic representations being automatically primed during comprehension, which "percolates" across levels, such that alignment at one linguistic level leads to alignment at other levels as well. This claim has been supported by evidence showing that lexical and semantic alignment "boost" syntactic alignment (Branigan et al., 2000; Cleland & Pickering, 2003; Mahowald et al., 2016). However, so far, there is little evidence that this would generalize to alignment across modalities. While one study reported that various verbal and non-verbal channels show reliable covariation (Louwerse et al., 2012), more fine-grained studies of lexical and gestural alignment found no correlation between alignment in the two modalities (Oben & Brône, 2016) and revealed that when gestures do not match a discourse context, they are unlikely to be copied, yielding alignment on the lexical level only (Mol et al., 2012).

In sum, prior qualitative work has demonstrated *how* alignment is employed as a resource for collaborative referring, with quantitative studies providing mixed evidence for *how frequently* lexical and gestural alignment (co-)occur. The two modalities can be recruited flexibly – yielding unimodal or multimodal alignment – which appears to be governed by the interactional needs at hand.

Modality and symbol creation

Talking about novel objects or concepts without conventionalized names brings along specific interactional challenges. If modality and alignment are indeed deployed flexibly to suit communicative demands (as previous work suggests; Chui, 2014; Holler & Wilkin, 2011; Mol et al., 2012; Oben & Brône, 2016), then the pressures of emergence contexts might invoke a preference for alignment in one particular modality, or they could call for the combination of alignment in both modalities. Though alignment has been studied in interactive tasks involving unfamiliar configurations (e.g., Fusaroli et al., 2012; Garrod & Anderson, 1987) or novel objects (e.g., Holler & Wilkin, 2011), we are not aware of any studies quantitatively investigating both lexical *and* gestural alignment in such settings. So, to derive hypotheses on the extent to which lexical and gestural alignment might be jointly or separately recruited when referring to novel referents, we turn to studies on language emergence and language development. Though not specifically targeting the phenomenon of alignment and its role in the process of shared symbol creation, this work is useful for its focus on contexts where conventional symbols are not yet established or acquired.

What the field of language emergence and language development have in common is the wealth of evidence for the importance of the manual modality. Children use gestures to refer to objects before they learn to produce words for those objects (Iverson & Goldin-Meadow, 2005) and have been shown to convey abstract concepts through gesture when they cannot yet do so in speech (Perry et al., 1988). Adults, too, employ the gestural modality as an effective means of communication when verbal labels are missing. In referential tasks, people have been shown to communicate more effectively when they use only gestures compared to only non-linguistic vocalizations (Fay et al., 2013, 2014; Macuch Silva et al., 2020; Zlatev et al., 2017), and more efficiently when they use multimodal symbols compared to either gestures or vocalizations alone (Macuch Silva et al., 2020; but see Fay et al., 2014 where there was no advantage for multimodal over gesture-only communication). Gestural and multimodal symbols probably offer such communicative benefits because of their versatility in establishing transparent form-meaning mappings: gestures can be used to visually depict object attributes, spatial relationships, actions, and motions. Through its iconicity and indexicality, gesture lends itself well for the production of motivated signs (i.e., signs that are linked to meaning by structural resemblance or by natural association; Fay et al., 2013; see also, Perniss & Vigliocco, 2014).

The iconic and indexical potential of gesture is one reason that gesture (alongside speech) is thought to play an important role in the initial stages of language evolution (e.g., Levinson & Holler, 2014; Sterelny, 2012; though there are "speech-first" accounts of language evolution too; Cheney & Seyfarth, 2005; MacNeilage, 2008; Mithen, 2005). Fay et al. (2013) argue that gesture is an effective means to bootstrap a communication system: "grounding a basic set of shared meanings in this way, during the very earliest stages of language, could then pave the way for the further expansion of the lexicon" (p. 1365).

In sum, when people align their behavior, they are likely to do so in one or multiple modalities depending on the communicative demands. The communicative demands of symbol creation settings (i.e., settings where conventionalized referring expressions are not yet established) appear to call for the use of gestural and/or multimodal symbols, though little is known about the use of alignment of those symbols during social interaction. Here, we aim to take the next step: we examine the interplay of lexical and gestural alignment in referring to novel referents. Combining quantitative and qualitative analyses, we chart the emergence of alignment in relation to modality and capture the interactional dynamics of how unimodal and multimodal alignment are employed for communicative purposes.

Present study

We aim to investigate how frequently, when and how alignment of co-speech gestures and lexical choice emerge when converging on shared symbols for novel referents. To do so, we use a multimodal corpus of interactions where people negotiate referring expressions for novel objects, which allows for (i) systematic quantitative observations of lexical and gestural alignment for particular referents, and (ii) qualitative inspection of the communicative environment in which alignment naturally unfolds.

We used a referential communication task in which participants used speech and gesture freely as they took turns to describe and find images of novel 3D objects, over six consecutive rounds. We also asked participants to individually name the objects both before and after the interaction. This set-up enables us to investigate:

(i) the extent to which participants managed to create shared symbols for the novel objects;
(ii) *how frequently* alignment emerges in the lexical modality only, the gestural modality only, or in both modalities in the interaction;
(iii) *when* alignment emerges in the lexical and gestural modality in the interaction;
(iv) *how* the different alignment patterns – independent, simultaneous, or successive emergence of lexical and gestural alignment – are functionally deployed to effectively refer to novel referents.

We expect participants to establish referential conventions during the interaction. The pre- and post-interaction naming of the objects serves as a rough proxy for the creation of such shared symbols, and so we hypothesize that participants will use more similar names to label the objects after the interaction, compared to before the interaction (prediction 1). Given that the interaction in our task is multimodal, we expect participants to recruit both lexical and gestural alignment as interactional resources for collaborative referring. Since participants share a spoken language, we expect that participants will work toward alignment on lexical choice, as shared lexical symbols are arguably more robust and efficient compared to relatively unconventionalized co-speech gestures. So, we predict that multimodal alignment and lexical alignment will emerge more frequently than gestural alignment alone (prediction 2).

We do not have a specific hypothesis as to whether alignment in both modalities will be more or less frequent than alignment in lexical choice only. Multimodal alignment might be expected based on psycholinguistic research showing that speech and gesture are produced and comprehended in an integrated way (Kelly et al., 2010; Kita & Özyürek, 2003; McNeill, 1992), yielding benefits of multimodality for message comprehension (Hostetter, 2011); as well as based on work underlining the affordances of the gestural modality for referential communication (especially in language emergence contexts, cf., section 1.2). However, this might not necessarily result in frequent use of multimodal alignment in this task, because people differ substantially in the amount of gestures they naturally produce, which has consequences for the opportunities for gestural alignment (Özer & Göksun, 2020).

Our third prediction concerns the temporal relation between lexical and gestural alignment in cases where multimodal alignment is deployed. Alignment can emerge in both modalities at the same time (simultaneous emergence), or alignment in one modality could precede alignment in the other modality (successive emergence). Based on prior work in the domains of language emergence and development, we expect that gestural alignment will either emerge together with lexical alignment or precede it, and we expect that least frequently of all, gestural alignment follows lexical alignment (prediction 3).

Quantitative analyses necessarily abstract away from important details of how alignment is interactionally achieved in the turn-by-turn context of conversational sequences. We attend to these details through qualitative, sequential analysis of the communicative environments in which lexical, gestural, and/or multimodal alignment naturally unfold. This ensures empirical grounding for the quantitative analyses and sheds light on how modality in alignment is employed to establish joint reference to novel referents.

Methods

Dataset

The current study is based on data collected within a larger research project aimed at investigating various kinds of cross-speaker alignment. For this project, participants performed a referential communication task, similar to the classic Tangram task (Clark & Wilkes-Gibbs, 1986; Holler & Wilkin, 2011), with images of 3D objects. Before and after this interactive task, participants individually named the objects: the naming task. For the current study, we draw on a subset of this dataset, by analyzing data from half of the dyads and half of the objects.

Participants

We analyzed data from 20 Dutch participants (11 women and 9 men, M_{age} = 22.9 years, $Range_{\text{age}}$ = 18–32 years). Prior to the task, the unacquainted participants were randomly grouped into dyads, resulting in 7 same-gender dyads (3 male dyads, 4 female dyads) and 3 mixed-gender dyads. The participants were recruited via the Radboud SONA participant pool system. Participants provided informed consent prior to starting the experiment and were paid for participation (12–16 euros, depending on total participation time). The study met the criteria of the blanket ethical approval for standard studies of the Commission for Human Research Arnhem-Nijmegen (DCCN CMO 2014/288).

Apparatus and materials

We used a set of 16 "Fribbles" (Figure 1 displays the 8 used in the analyses of the present study), illustrations of novel three-dimensional objects (based on Barry et al., 2014), designed in such a way as to ensure cross-participant and cross-dyadic variation in elicited names. During both the naming task and the interactive task, all 16 Fribbles were simultaneously presented on a gray background in a size

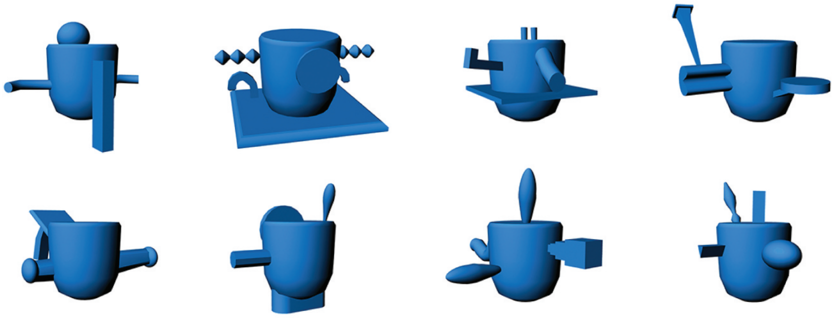


Figure 1. “Fribbles” that were used as stimuli; selection of 8 that were used for the analyses.

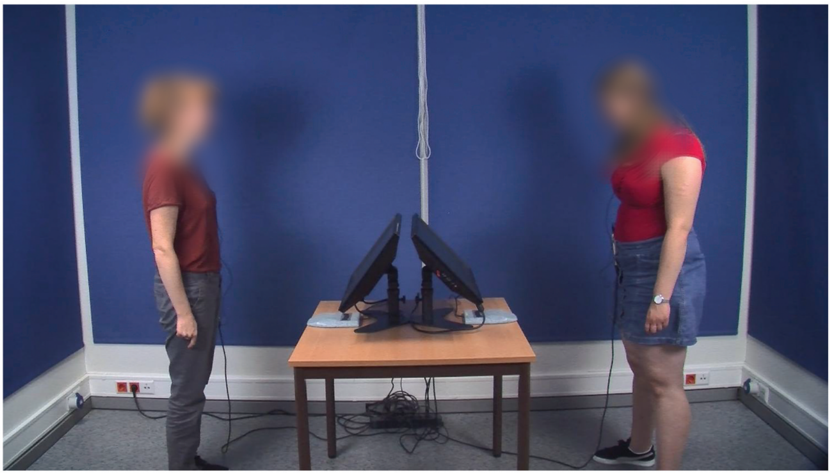


Figure 2. Set-up during interactive director-matcher task.

of about 4 × 4 cm per figure. The Fribbles were randomly distributed over 16 positions (forming rows of 5, 6, and 5 items, respectively). In the interactive, but not in the naming task, the Fribbles were labeled with letters for one participant, and numbers for the other (see Section “Procedure”). The naming task was conducted in two separate booths, where each participant was seated in front of a computer screen and used a keyboard to name the Fribbles. In the interactive task, participants were standing and faced each other (see Figure 2). Each had their own 24′ screen (BenQ XL2430T), slightly tilted so participants could easily view the screen and their partner, and positioned at hip height to ensure mutual visibility of upper torso and gesturing area. Each participant had a button box to move to the next trial. Verbal and nonverbal behavior was recorded using two head-mounted microphones (Samson QV) and three HD cameras (JVC GY-HM100/150).

Procedure

In the naming task, participants were asked to give a name or description of 1 up to 3 words for each image (i.e., the Fribbles) in such a way that their partner (the other participant) would be able to find it among the other images. Target Fribbles were indicated with a red rectangle, and participants could use “ENTER” to move to the next Fribble (the order was randomized across participants). During this task, participants knew that they would take part in a communicative task afterward, but they were not informed that this would involve the same images, nor that they would have to do the naming task again afterward. The naming task before the interaction took 5.41 minutes on average (range = 2.24–8.01 minutes).

The referential communication task consisted of six consecutive rounds, consisting of 16 trials each, with director and matcher roles alternating after each trial. In each trial, a single target Fribble was highlighted for the director by means of a red rectangle. Participants were instructed to work together in order to come to a shared understanding of what the target item is. The order in which the Fribbles were presented on the screen varied across the participants. To avoid confusion about the different orders, the Fribbles were labeled with numbers for one participant and letters for the other. Once the matcher was confident they identified the item described by the director, they said the corresponding positional label out loud and pressed a button to go to the next trial. Once all 16 trials had been completed, the Fribbles were shuffled and a next round started. The trial order was such that each participant took on the director role for a certain Fribble either in rounds 1, 3, and 5 or in rounds 2, 4, and 6. No time constraints were posed and the participants did not receive feedback about accuracy. Participants were told that they were "free to communicate in any way they wanted" (an instruction phrased to be agnostic about communicative modality, i.e., speech and/or gesture), and that their performance would be a joint achievement. The communicative task lasted for 24.92 minutes on average (range = 16.38–34.56 minutes).

After the interaction, participants again individually named the Fribbles, with the same instructions as before (the only change was an additional sentence stating that the name could be the same as before, but did not have to be). This took 1.89 minutes on average (range = 0.87–3.14 minutes).

Analysis

To assess the extent to which dyads had shared symbols for the Fribbles before and after the interaction, we computed the similarity of the names they provided in the naming task. We considered names to be similar when they consisted of the same base words. All words were first spell checked, lemmatized (i.e., inflected verbs changed into infinitives, plural and diminutive forms into singular nouns) and compounds were split if they were not standard Dutch words (verified with the online Van Dale dictionary). Naming similarity was computed by taking the cosine similarity of the participants' names (i.e., vectors of words), resulting in a score ranging from 0 (no similarity) to 1 (perfect similarity; cf., Duran et al., 2019 where the same measure is used for computing lexical alignment). For example, the comparison of "right round disk" and "disk horizontal right" resulted in a similarity score of 0.67.

Since the Fribbles are new to participants, the interactive task primarily involved talking about them in terms of subparts (each Fribble has about four distinctive subparts, while the "base" figure is the same, see Figure 1). We took these subparts as the primary target of possible alignment in gesture and/or speech, so they make up the main unit of analysis in this study. To keep the amount of hand-coded data manageable, here we analyze half of the target items (i.e., 8 out of 16 Fribbles, with a total of 34 subparts, see Figure 1). We arbitrarily selected which half to use, while ensuring that the dataset remained balanced (i.e., participants start as a director viz. matcher in the first round for four items each).

Transcription and coding of multimodal interaction

Transcription of speech and annotation of gestures was done in ELAN (version 5.8). Speech was segmented into Turn Constructional Units (TCU; Couper-Kuhlen & Selting, 2017; Schegloff, 2007) and orthographically transcribed based on the standard spelling conventions of Dutch. For co-speech gestures, only the stroke phase was annotated (i.e., the meaningful part of the gestural movement; Kendon, 2004; McNeill, 1992), for the left and right hand separately. Gestures were categorized into three types: 1) iconic gestures, which depict physical qualities of concrete referents or movements or

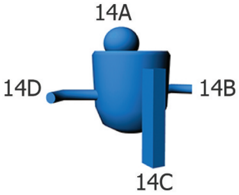


Figure 3. Example of Fribble subpart codes as used in coding protocols and transcripts.

actions related to those referents, 2) deictic gestures, or pointing gestures, and 3) other gestures, which were mostly beat gestures and interactive gestures. Only the first category (iconic gestures) was used in the analyses below.

For the iconic gestures, we coded which Fribble subpart(s) the gesture referred to, using a predefined coding protocol as illustrated for Fribble 14 in Figure 3. Gesture referents were coded based on the kinematics of the gesture together with the co-occurring speech and overall discourse context. Gestures can refer to one subpart (e.g., a curved hand as if holding a ball to depict 14A), or to more than one subpart simultaneously (e.g., using both arms alongside the body to represent 14B+14D).

Inter-rater reliability for gesture identification and gesture coding (gesture type and gesture referent) was moderate to high (for details of the inter-rater reliability analyses and results, see Appendix A).

Operationalization of alignment

Any notion of communicative alignment makes relevant an operationalization with respect to five dimensions: sequence, time, meaning, modality, and form (Rasenberg et al., 2020). Our research questions primarily concern the sequential and temporal patterning of alignment, so we impose no a priori restrictions on the dimensions of sequence or time (so two instances of similar behavior may count as aligned whether they occur within the same sequence or round, or at larger time spans across sequences and rounds). We fixate the phenomenon by focusing on the remaining three dimensions. For meaning, our criterion is referential alignment: we consider cross-speaker repetition of words or gestures to be a case of alignment only if they are used to refer to the same referent, and we exclude non-referential speech and gestures. So, if both participants use the word *egg* to refer to Fribble subpart 14A, this would count as lexical alignment, but not if one of them used it to describe another Fribble subpart. For modality, we look at alignment *within* modalities (comparing words with words and gestures with gestures), not across modalities. For form, finally, we use modality-specific criteria designed to yield a maximally commensurate measure of form similarity across modalities, as detailed in Appendix B. To summarize our criteria, we consider lexical choice to be aligned if there is at least one common word (after lemmatizing) that both participants use to refer to the same referent, and which is informative for distinguishing referents. We consider gestural behavior to be aligned if both participants use an iconic gesture to refer to the same referent. This is based on an explorative analysis showing that the majority of those gesture pairs overlap in one or more form features, even though exact copies are rare (see Appendix B for detailed results).

Quantitative and qualitative analyses of alignment

Our analyses were performed on the level of Fribble subparts ($N = 340$; 10 dyads * 34 subparts), where we first disregarded subparts that were never referred to (with speech or gesture) by either one or both members of a dyad, as by definition alignment would be impossible in those cases. For the remaining subparts ($n = 276$), we investigated whether dyads aligned lexically and/or gesturally in their referring expressions. For each case of alignment, we inspected when the “first element” (i.e., the initial word or gesture) and “second element” (i.e., the first time that word or gesture is used by the other speaker) were produced. We consider alignment to have emerged at the moment the second element is produced. Note that temporal distance between the respective elements can vary greatly (e.g., they might occur in adjacent turns within a trial, but also in different rounds of the interaction). Since we are interested in

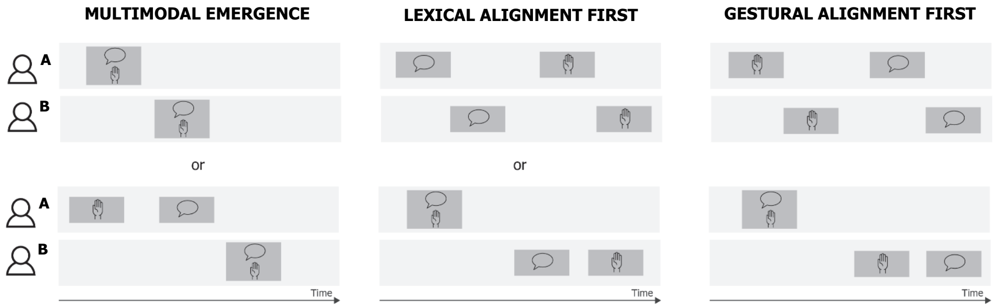


Figure 4. Examples of how temporal order of emergence is categorized when alignment is achieved both lexically and gesturally. Speech balloons icons are used for speech, and hand icons for co-speech gestures. Grey rectangles represent TCUs.

the emergence of alignment, we only coded the first occurrence of alignment for a given modality. To exemplify: once we found the emergence of lexical alignment, we did not code later re-occurrences of the aligned-upon verbal expression, nor did we check whether the dyad aligned on a different set of words later on.

In sum, for each Fribble subpart that both members of a dyad referred to, we noted in which modality/modalities alignment emerged (NO ALIGNMENT, LEXICAL ONLY, GESTURAL ONLY, or MULTIMODAL), as well as when it emerged (i.e., in which of the six rounds of the interaction the second aligning element was produced). When multimodal alignment emerged for a Fribble subpart, we grouped it into one of three categories: MULTIMODAL EMERGENCE, LEXICAL FIRST, or GESTURAL FIRST. We regarded a case as MULTIMODAL EMERGENCE when the second element of both lexical and gestural alignment was produced in the same TCU. We coded a case as LEXICAL FIRST if lexical alignment had emerged earlier than gestural alignment, that is, when the second element of the lexically aligned pair occurred in an earlier TCU than the second element of the gesturally aligned pair; and vice versa for the category GESTURAL FIRST (see Figure 4). The cases thus identified formed the dataset for which quantitative and qualitative analysis were conducted.

To analyze shared symbol creation, we used a paired samples Wilcoxon signed-rank test to assess whether naming similarity was higher after compared to before the interaction (prediction 1). To compare the frequencies of the alignment categories and orders of emergence, we used intercept-only mixed effects models with random intercepts for dyads and subparts (unless otherwise specified). These were binomial models, where specific categories were coded as 0 versus 1 to test the comparisons as specified in the hypotheses (predictions 2 and 3). Finally, we used two sample, two-sided Kolmogorov-Smirnov tests to exploratively compare categories in terms of their distributions of time of emergence (i.e., in which round of the interaction alignment emerged).

For the qualitative analysis, we used observational methods from interactional linguistics and conversation analysis (Clift, 2016; Couper-Kuhlen & Selting, 2017) to make visible the interactional work that participants accomplish with alignment. This allows us to study the sequential and formal properties of multimodal alignment as it emerges in interaction, enriching our understanding of the quantitative patterns.

Results

Shared symbols in the naming task

To find out to what extent dyads created shared symbols for the Fribbles, we compared how similar the names (consisting of 1 to 3 words) were that members of a dyad used to label a Fribble, both before the interaction (*pre*) and after the interaction (*post*); see Figure 5, panel A. As expected, we found that the naming similarity scores increased from *pre* ($M = 0.07$, $Median = 0$) to *post* ($M = 0.46$, $Median = 0.41$). A paired samples Wilcoxon signed-rank test indicated that this difference was statistically significant ($Z = 0.70$, $p < .001$).

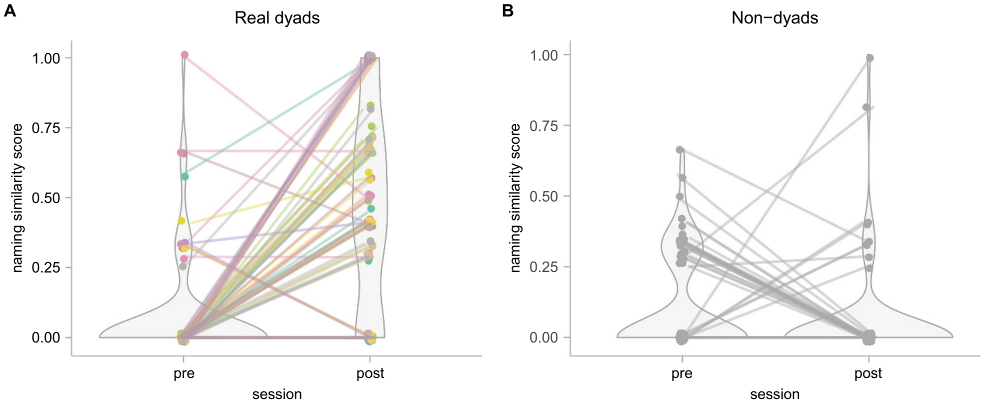


Figure 5. Distribution of naming similarity scores (i.e., cosine similarity of overlapping words in the names provided by participant A and B of a dyad for a particular Fribble), before (pre) and after (post) the interaction. Results from real dyads (panel **A**) are contrasted with those from non-dyads (i.e., pairs which did not interact with each other; panel **B**). Dots represent individual datapoints ($N = 80$); colors represent dyads ($N = 10$).

By itself this leaves unclear whether the increased naming similarity is contingent on the history of the interaction, or simply a result of spending time with the stimuli and repeatedly formulating references over six rounds. To tease these options apart, we compared the scores of "real dyads" with those of "non-dyads" (i.e., people who did not interact with each other; see Figure 5, panel B). We computed the non-dyad scores with a simple shifting function, where all names from participants B were paired with the names from participant A from the next dyad, while keeping Fribble and Session (*pre/post*) constant. In contrast to the real dyads, for the non-dyads, there is no systematic improvement from *pre* ($M = 0.11$, $Median = 0$) to *post* ($M = 0.05$, $Median = 0$; $p = .06$). This allows the inference that symbol creation was indeed contingent on dyadic interaction.

Remarkably, even for real dyads there are quite some name pairs with zero similarity *post* interaction ($n = 21$). Further investigation revealed that these were often cases where the two members of a dyad labeled different subparts of a Fribble. For example, participant A's name referred to the orientation with respect to one subpart ("stands on rectangle"), while participant B's name referred to another subpart ("spoon top right"). Conversely, names with naming similarity scores of 1 ($n = 12$) were usually labels for one specific subpart (e.g., "chimney") or a more holistic name for the whole Fribble (e.g., "rabbit").

Though the *post* naming similarity scores might be expected to follow from the degree of alignment in the interaction, an explorative investigation yielded no evident relationship between the two (see Appendix C). This is unsurprising given the fact that the naming task elicits short written forms at the level of whole Fribbles, whereas for our measure of alignment we focused on the emergence of alignment in both speech and gesture, and at the level of the subparts, creating many opportunities for differences in selection and construal. We will get back to this in the Discussion.

Prevalence of alignment in the interactive task

Task performance was high, with matchers selecting the correct target Fribble in 99.8% of the trials. Alignment was highly frequent in the task: on average across dyads, alignment emerged in at least one modality at some point in the interaction for 92% of Fribble subparts that had been (lexically and/or gesturally) referred to by both members of a dyad. For the subparts where alignment emerged ($n = 255$), 56% involved multimodal alignment, 38% lexical alignment only, and 6% gestural alignment only (see Figure 6). As predicted, gestural alignment only occurred less frequently than multimodal alignment ($\beta = 2.53$, $SE = 0.49$, $z = 5.53$, $p < .001$)[1] and lexical alignment only ($\beta = 2.08$, $SE = 1.05$, $z = 1.97$, $p = .048$).[2]

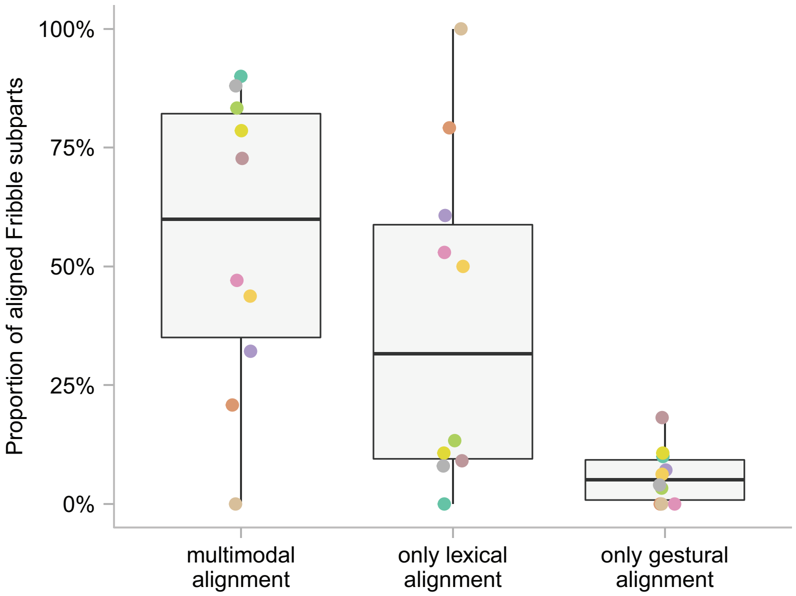


Figure 6. Average proportion of Fribble subparts (that have been referred to by both participants of a dyad) for which alignment occurred, by modality. Colored dots represent dyads ($N = 10$).

Temporal distribution of unimodal and multimodal alignment

In answering *when* lexical and gestural alignment are deployed to refer to novel referents, we first compare unimodal with multimodal alignment. Alignment tended to emerge early in the interaction: for 80% of the subparts for which alignment was achieved, it emerged in the first or second round (Figure 7). Note that emergence in the second round is more common than in the first. This is to be expected because director/matcher roles switched over trials; directors usually (lexically and/or gesturally) described the Fribbles extensively in the first round (while the contributions from matchers varied), which was then "aligned to" in the second round by the other participant when taking up the role of director for that Fribble.

Early emergence was especially prevalent for multimodal alignment. The first instance of alignment emerged in the first or second round in 92% of the multimodally aligned subparts (Figure 7, panel A). Emergence in rounds 1 or 2 occurred less frequently for unimodal alignment, with 71% for lexical only and 50% for gestural only (Figure 7, panels B and C). Kolmogorov-Smirnov tests revealed that the distribution of time of emergence was different for the category multimodal alignment when compared to lexical alignment only ($p = .018$) and gestural alignment only ($p = .013$); the distributions of the latter two categories did not differ significantly ($p = .560$).[3]

Order of emergence in multimodal alignment

For cases of multimodal alignment, we investigated whether lexical and gestural alignment emerged simultaneously, or whether alignment in one modality preceded alignment in the other modality. For the subparts where alignment emerged in both modalities ($n = 148$), we found that emergence was simultaneous in 51% of cases; gestural preceded lexical alignment in 28% of cases; and lexical preceded gestural alignment in 21% of cases (Figure 8, panel A). As predicted, simultaneous multimodal emergence occurred more frequently than lexical alignment

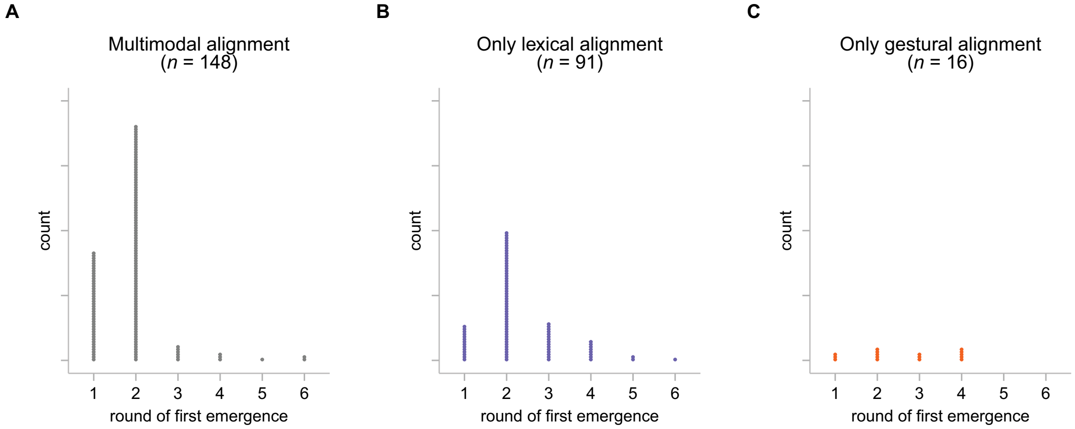


Figure 7. Distribution of rounds of the interaction in which alignment first emerged. For multimodal alignment this represents the time point of the first instance of alignment in either modality (see Figure 8 for details).

first (β = 0.94, *SE* = 0.34, *z* = 2.74, *p* = .006). But contrary to our hypothesis, we found no evidence for a difference between the frequency of lexical alignment first and gestural alignment first (β = 0.33, *SE* = 0.30, *z* = 1.10, *p* = .274).[4,5]

To explore the relation between order of emergence and time of emergence, we compared the temporal distributions of the first instance of alignment for the three categories (see the density plots in Figure 8, panels B-D). The Kolmogorov-Smirnov tests revealed no differences between the three categories (all *p* > .05).

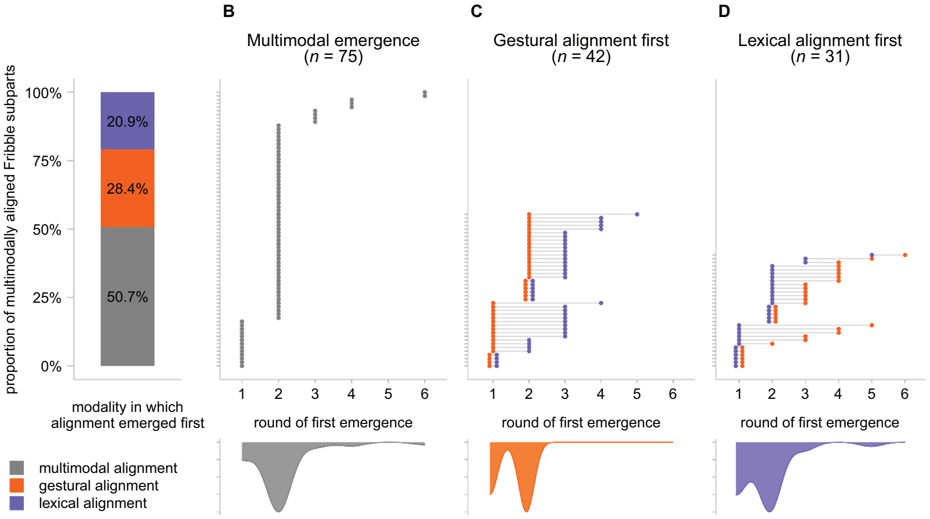


Figure 8. Temporal order of the emergence of lexical and gestural alignment. Panel **A** shows that simultaneous, multimodal emergence is most frequent, followed by gestural alignment preceding lexical alignment and lexical alignment preceding gestural alignment. The dumbbell plots in panels **B**-**D** display in which rounds of the interaction lexical and gestural components of multimodal alignment emerged (ticks on y-axis represent individual datapoints; i.e., Fribble subparts). For example, in panel **C**, the very top row shows that for one particular Fribble subpart, gestural alignment emerged in round 2, followed by lexical alignment in round 5. The bottom plots are density plots corresponding to the (first) dots of the dumbbell plots above.

Multimodal alignment: qualitative analyses

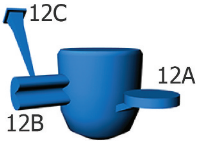
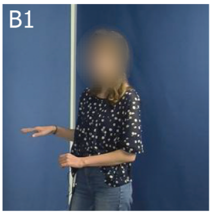
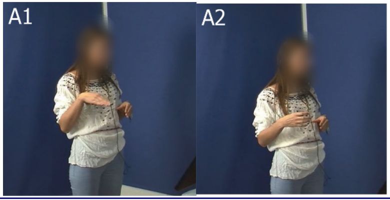
With the quantitative evidence in hand, we are in a position to consider qualitative evidence for *how* lexical and gestural alignment are recruited as interactional resources. Multimodal emergence of alignment (i.e., simultaneous emergence of lexical and gestural alignment) most often consisted of cases where lexical and gestural alignment went "hand in hand," where a particular composite utterance (e.g., "ball" + ball gesture) was repeated as a whole by the other speaker (n = 67). Transcript 1 shows a representative case of how alignment emerged multimodally in the interaction. In all transcripts, "A" and "B" refer to participants A (standing on the left side) and B (on the right), and the underlined speech temporally overlaps with the gesture strokes depicted in the video still with the corresponding subscript (cf., Mondada, 2018).

Here, the director (A) confirms the matcher's question about subpart 12A through verbal and gestural repetition (i.e., repetition of the noun "plateau" and the accompanying gesture), with meaningful variation to provide further information. She adds the adjective "circular," which is also expressed gesturally by adding a circular motion to gesture A1 (which otherwise looks similar to the matcher's gesture B1). So the director refashions the presented referential expression through what has been called "expansion," though rather than a mere verbal process (as in the original account by Clark & Wilkes-Gibbs, 1986), here it is done in both speech and gesture.

Besides cases where lexical and gestural alignment emerge "hand in hand," there was a less frequent pattern of multimodal emergence (n = 8), where the first word and gesture were *not* produced in a single speech turn, while the repeated word and gesture were (see visualization of this distinction in Figure 4). For example, one speaker introduced the lexical choice "zeppelin" in an initial TCU, which was followed by a gestural depiction with different co-expressive speech in the next TCU. Yet, later on they were produced together as one composite utterance by the other speaker ("zeppelin" + gesture), yielding simultaneous multimodal alignment.

When multimodal alignment is not simultaneous, there appear to be two types of temporal patterns of successive occurrence (Figure 8, panels C and D). First, lexical and gestural alignment can closely succeed each other, where both emerge within the same round. We find this pattern in both directions: sometimes lexical alignment emerged first, followed by gestural alignment; and vice versa, gestural alignment first, shortly followed by lexical alignment. Such close successions of lexical/gestural alignment only occurred in the first or second round of the interaction. Alternatively, alignment could emerge at larger sequential and temporal distances (e.g., gestural alignment in round 1, followed

Transcript 1. Simultaneous emergence of gestural and lexical alignment when "expanding" a referential expression.

	round 1	B (matcher):	ja en met zo'n **plateautje** aan de rechterkant $_{B1}$? *yes and with a **plateau** (like this) on the right side$_{B1}$?*	
		A:	ja een soort uh cirkelachtig $_{A1}$ **plateautje** $_{A2}$ inderdaad *yes a sort of uh circular$_{A1}$ **plateau**$_{A2}$ indeed*	

$_{B1:}$ right-handed gesture depicting the horizontal orientation and relative position of 12A to the base shape; the flat palm-down hand makes small lateral movements.

$_{A1:}$ right-handed gesture depicting 12A (similar to $_{B1}$), with a circular motion depicting the shape.

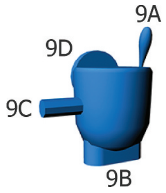
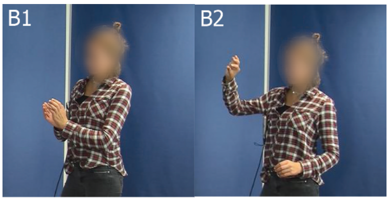
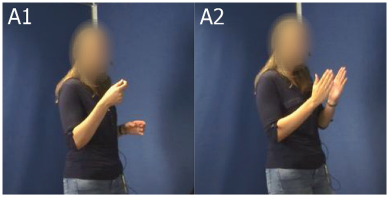
$_{A2:}$ right-handed gesture depicting 12A, with a curved handshape depicting the shape.

by lexical alignment in round 3), which sometimes involves very late emergence for one modality (even as late as round 6). These two types of patterns appear to be qualitatively different from each other, and will be discussed in turn.

Transcript 2 provides an example of the pattern in which the emergence of gestural alignment is followed by the emergence of lexical alignment in relatively close succession. In round 1, participant B referred to subpart 9A with the words "cone" and "upside down," along with depictive gestures representing the same subpart. In round 2, participant A runs into trouble verbally describing the subpart, and produces a disfluent utterance supported by two depictive gestures that resemble both of participant B's earlier gestures. The emerging gestural alignment appears to be used here in search of lexical convergence. Participant B gazes at participant A's gestures and then suggests a lexical completion for participant A's utterance ("cone?"), which participant A accepts while seeking and receiving further clarification of the fuller lexical formulation ("a cone upside down"), establishing lexical alignment. The interactional work done by the gestures appears to support a word search and is likely aided by their visible similarity.

We see the reverse, with lexical alignment coming first, in Transcript 3. Here both participants use "disk" to refer to 12A: in the first round produced by A as the director without a gesture, and in the second round by B as the director with a gesture. However, the combination of B's noun phrase ("horizontal disk") and gesture (representing the horizontal orientation of the disk with a sharp lateral movement) is treated as inconclusive by A, who seeks to clarify the *shape* of the subpart. This is done, much like in Transcript 1, by presenting a modified version of both the noun phrase ("round disk" instead of "horizontal disk") and the gesture (as if molding a disk, with a curved handshape), establishing gestural alignment in the process. Although a partial form of lexical alignment was

Transcript 2. Gestural alignment preceding lexical alignment in search of lexical convergence.

	round 1	B (director):	en dan boven steekt dus laat maar zeggen $_{B1}$ zo'n op- ja **op de kop** zo'n **kegel** uit $_{B2}$ *and then on top stick so let's say $_{B1}$ a up-yes **upside down** (this kind of) a **cone out** $_{B2}$*	
	round 2	A (director):	en rechts bovenop de ronde, op, bovenop de hoofdvorm heb je een soort van $_{A1}$ (.) ja $_{A2}$ *and right on top of the round, on, on top of the main shape you have a sort of$_{A1}$ (.) yes$_{A2}$*	
		B:	kegel? *cone?*	
		A:	uh ja uh hoe noem je zoiets? een uh *uh yes how do you call something like that? a uh*	
		B:	kegel op de kop *cone upside down*	
		A:	ja een **kegel op de kop** inderdaad *yes a **cone upside down** indeed*	

$_{B1:}$ two-handed gesture depicting the shape of 9A; static gesture with the wrists held together and the curved palms slightly apart.

$_{B2:}$ right-handed gesture depicting 9A; the index finger and thumb are held slightly apart (illustrating the width of the subpart), while making a single upward (slightly diagonal) movement, depicting the orientation of the subpart.

$_{A1:}$ right-handed gesture depicting 9A (similar to $_{B2}$); the index finger and thumb are held slightly apart, while making an up-and-down movement.

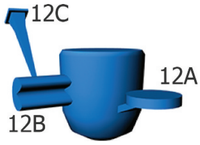
$_{A2:}$ two-handed gesture depicting 9A (similar to $_{B1}$): the hands start out put against each other, then move upward with the palms slightly apart, and end with the fingertips touching each other.

established at the start of round 2 (where participants align on the noun ("disk"), but not on the adjective ("round" versus "horizontal")), the subsequent lexical and gestural refinements serve to further disambiguate and calibrate the emerging multimodal conceptual pact.

Transcripts 2 and 3 demonstrated how alignment in the two modalities emerge in close succession early on in the interaction, working together to establish mutual understanding. We now turn to the patterns of more distant emergence, starting with the category gestural alignment first. Participants frequently use gestures to establish joint reference early on the interaction (with gestural alignment emerging in round 1 or 2, while the lexical references are not yet aligned, or rather underspecified, e.g., "protrusion"), which is later on followed by lexical alignment (e.g., "horn" in round 4). With respect to the category lexical alignment first, participants at times appear to resort to gestures later on in the interaction to deal with interactional trouble, such as to further calibrate a (somewhat underspecified or partial) lexical pact (much like in Transcript 3) or when they appear to have trouble retrieving a lexical item, as shown in Transcript 4.

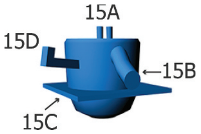
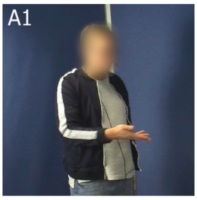
Though lexical alignment emerged in round 5, in round 6, participant A's description of this Fribble runs into disfluency ("and the and the and the"), foreshadowing trouble in retrieving a lexical item. He finally produces a lexical item ("plane") that is different from the one they aligned on before, but does so together with a gestural depiction of 15C, using gestures that are similar to those produced much earlier by B (in rounds 1 and 3). So, a similarity in gestural representation is used to restore collaborative reference. The use of gesture in an environment of disfluent speech is similar to what we saw in Transcript 2, and underlines the flexible way in which language users shift the division of labor across modalities. Two non-exclusive ways to interpret the use of gesture here are that gesture helps lexical retrieval and/or that gesture is used as compensation for the "broken" lexical pact.

Transcript 3. Lexical alignment followed by gestural alignment for calibrating a conceptual pact.

	round 1	A (director):	uh deze heeft aan de rechterkant een platte ronde **schijf** en aan de linkerkant heb je een uitsteeksel met daar bovenop nog zo'n heel langwerpige [zo'n toeterding *uh this one has on the right side a flat round **disk** and on the left side you have a projection with on top of that another like very elongated [such a horn thing*	
		B:	[ja G *[yes G*	
	round 2	B (director):	dit is die uh de beker waarvan er uh een horizontale **schijf** $_{B1}$ rechts zit en dan links zit nog een uitsteeksel met zo'n hele lange ja kegel *this is that uh the cup of which uh one horizontal **disk**$_{B1}$ is on the right and then on the left there is another projection with such a very long uh conea*	
		A:	[oh ja *oh yes*	
		B:	[zo'n spijl erbovenuit *[such a bar above*	
		A:	en aan de rechterkant zo'n ronde **schijf** $_{A1}$ toch? *and on the right side such a round **disk**$_{A1}$ right?*	
		B:	ja gewoon die plat staat ja *yes just which is flat yes*	
		A:	ja 15 *yes 15*	

$_{B1:}$ left-handed gesture where the hand models 12A, with a sharp lateral movement marking the horizontal orientation.
$_{A1:}$ left-handed gesture with a curved handshape depicting the shape of 12A.

Transcript 4. Gestural alignment in an environment of lexical disfluency.

	round 6	A (director):	uh dit is de glijbaan, de pijp en de en het en het vlak$_{A1}$ dat er doorheen zit *uh this is the slide, the pipe and the and the and the plane$_{A1}$ which is through it*	
		B:	yes dat is L *yes that is L*	

$_{A1:}$ right-handed gesture depicting 15C; flat hand palm-up, making a lateral movement depicting the horizontal orientation.

Unimodal alignment: qualitative analyses

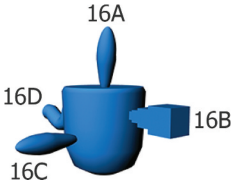
While multimodal alignment was prevalent and emerged early in the interactive task, both lexical and gestural alignment separately also warrant our analytical attention, starting with lexical alignment only (the most common case after multimodal alignment).

Lexical alignment was most likely to emerge in round 2. It is useful to look more closely at the interactional work alignment is doing in such cases. We found that often a director produced a particular noun phrase in round 1, which was reused by their partner when taking on the role of director in round 2. But this reuse was rarely straightforward repetition and typically involved some modification or expansion. Consider Transcript 5.

In round 1, the matcher appears to have found the target Fribble (as suggested by an inbreath and a stretched change of state token "o:h"), and subsequently describes several other subparts of the Fribble to verify her selection. After describing subpart 16B as a "square nose," she goes on to describe 16D, but is interrupted by the director who completes her sentence with *slurfje, staartje* "[elephant's] trunk, tail," which A confirms by saying "yes." This double-barreled candidate description (casting part 16D as a small trunk or tail) provides source material for a conceptual pact, but does not yet commit to a single conceptualization; indeed, the two candidate nouns imply opposite animal parts. In the next round, participant A (now director) reuses B's word "trunk" in her description. The immediate result of this case of lexical alignment is to commit to one particular conceptualization, which is taken up without further problems by B. Though this example came from a dyad where both participants gestured regularly, it shows that sometimes lexical alignment can be sufficient for the task at hand.

Turning to the category of gestural alignment only, even if this is relatively rare, two salient patterns emerged in the data. The first one is where gestural alignment emerges early on for subparts that may be hard to capture in speech, as shown in Transcript 6.

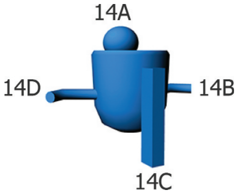
Transcript 5. Lexical alignment for calibrating a conceptual pact.

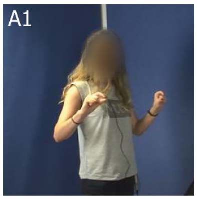
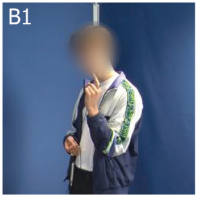
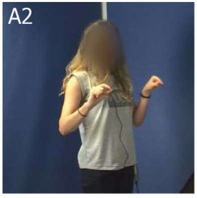
	round 1	A (matcher):	((inbreath)) o:h ja maar hij heeft ook één zo'n uh vierkante neus *o:h yes but he also has one like uh square nose*
		B:	ja *yes*
		A:	en nog een soort *and also a sort of*
		B:	**slurfje**, staartje ***trunk**, tail*
		A:	ja *yes*
	round 2	A (director):	met een vierkante schroef als neus en een **slurfje** aan de achterkant *with a square screw as nose and a **trunk** on the back*

Here, A and B refer to 14D, both verbally and gesturally. Both start speaking in overlap, with A resolving the overlap by withholding completion of the spoken turn while launching into a depictive gesture. B's turn is completed in the clear with an alternative gestural depiction occupying the slot of the noun (Clark, 2016). This composite utterance is treated as sufficient by A, as seen by her spoken confirmation and another gesture produced with her left hand (which she still had in the air, i.e., in a post-stroke hold). However, she somewhat changes the gesture's handshape and motion (now showing more resemblance to B's gesture), as if to say "what you just gestured is the same as what I was gesturing about." With the spoken utterances conveying only limited information, the dyad appears to rely heavily on coordinating their gestures to achieve collaborative reference.

The second pattern of gestural only alignment is where speakers resort to gestures for a particular referent throughout (most) of the interaction in a way that compensates for the lack of lexical alignment on that referent. Consider Transcript 7: throughout the interaction, the two speakers of a dyad used different nouns ("lumps" versus "spheres") to refer to subparts 10B+10F. While A has produced accompanying gestures in rounds 1 and 3, participant B produces a similar gesture for the first time as late as round 4. The sequential environment in which this happens is telling. After B's initial verbal description in round 4, A produces a soft verbal repetition of part of the formulation ("arms with . . . ") while visibly scanning the array of Fribbles on her screen. This display of trouble is followed by an upgraded formulation on the part of B, who now produces a multimodal utterance that is both more lexically specific ("two arms with a sphere attached") and features a two-handed gestural depiction of the spheres time-aligned with "sphere". So, where a mere lexical formulation proved insufficient for A, the dyad resorted to the gestural modality to establish collaborative reference, and continued to rely on the gestural depiction (in the absence of lexical alignment) in rounds 5 and 6 as well.

Transcript 6. Gestural alignment as a substitute for speech.



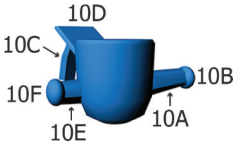
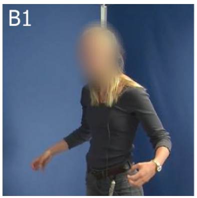
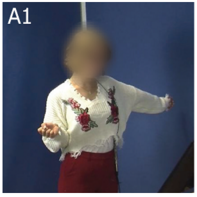
round 1	A (director):	[en aan de linkerkant$_{A1}$ *[and on the left side $_{A1}$*	
	B:	[en links zit nog een soort zo gebogen $_{B1}$ *[and on the left there is also a sort of bended$_{B1}$*	
	A:	ja $_{A2}$ *yes$_{A2}$*	
	B:	ja *yes*	
	A:	ja *yes*	

A1: left-handed gesture depicting subpart 14D; the index finger and thumb are held slightly apart (illustrating the width of the subpart), while making small sideward movements. The right-handed gesture is a post-stroke hold (depicting 14B, which is irrelevant for current purposes).

B1: left-handed gesture depicting subpart 14D; a curve is traced with the extended index-finger.

A2: left-handed gesture depicting subpart 14D; the index finger is slightly extended in a single sideward movement.

Transcript 7. Gestural alignment for reestablishing collaborative reference under uncertainty.

	round 4	B (director):	dit is die met uh de armen met bolletjes eraan *this is the one with uh the arms with spheres attached*	
		B:	en achter een uh ding nog *and behind a uh another thing*	
		A:	^{0}armen met^0 ((visibly searches on screen)) ^{0}*arms with*^{0}	
		B:	twee armen met een [bol $_{B1}$ eraan en dan *two arms with a [sphere*$_{B1}$ *attached and then*	
		A:	[oh die zo $_{A1}$ *[oh the one like (this)*$_{A1}$	
		B:	ja *yes*	

$_{B1:}$ two-handed gesture where curved handshapes depict the round shapes of subparts 10B+10F, somewhat away from the body thereby depicting the subparts' positions relative to the base shape.

$_{A1:}$ two-handed gesture where clenched fists model subparts 10B+10F, right extended arm models 10A and left tucked-in arm models 10E.

What unites both of these patterns of gestural-only alignment is that they rely on the visuo-spatial affordances of the gestural modality to achieve joint reference by iconically depicting aspects of a referent, either because it is hard to capture in speech, or because a spoken formulation turned out hard to interpret.

Discussion

Quantitative findings

With the present study we aimed to reveal how frequently, when and how lexical and gestural alignment emerge when creating shared symbols for novel referents. First of all, our results confirm that symbol creation took place over the course of the interaction, as we found that the names that participants used to label the novel objects were more similar to each other after compared to before the interaction (prediction 1 supported). As for the interactions, we found that alignment was very frequent overall: for 92% of the novel referent subparts that dyads referred to, some form of alignment occurred at some point in the interaction, with multimodal and lexical alignment being more frequent than gestural alignment only (prediction 2 supported). We found a distinctive pattern for multimodal alignment: it was both more frequent than gestural alignment only and tended to emerge earlier in the interaction compared to both lexical and gestural alignment only. For those cases of multimodal alignment, we found mixed support for prediction 3: emergence of alignment in both modalities simultaneously was more frequent than successive emergence (i.e., lexical alignment preceding gestural alignment or vice versa), but contrary to our expectations, the two types of successive emergence (gestural alignment preceding alignment, and lexical alignment preceding gestural alignment) were equally frequent.

The prevalence of alignment in our study corroborates the notion that alignment plays an important role in collaborative referring (Brennan & Clark, 1996; Fay et al., 2014, 2018; Holler & Wilkin, 2011; Reitter & Moore, 2014).[6] We found that lexical and gestural alignment can be deployed flexibly: they can occur in tandem as well as independently, which is in line with earlier work showing no systematic relation between the two (Oben & Brône, 2016), and qualitative reports on various combinations of lexical and gestural alignment (Chui, 2014; Holler & Wilkin, 2011; Bertrand et al., 2013). Yet, multimodal alignment was clearly favored. This finding relates to psycholinguistic work on multimodal communication in two ways. First, given that speech and gesture are integrated during both production and comprehension (Kelly et al., 2010; Kita & Özyürek, 2003; McNeill, 1992), multimodal alignment may be the result of cross-participant repetition of the composite utterance as a whole. Second, since receivers have been shown to benefit from multimodality in message comprehension (Hostetter, 2011), participants could have relied mostly on multimodal, rather than unimodal alignment, to ensure more robust communication in this task.

The prevalence of multimodal alignment also ties in with the previously reported efficiency advantage for multimodal signals in the field of experimental semiotics (Macuch Silva et al., 2020), and with accounts of multimodal origins of language (Levinson & Holler, 2014; Perlman, 2017; Zlatev et al., 2017). Our study complements this prior work by showing that when people cannot rely on conventionalized referring expressions, multimodality is not only a useful property of communicative signals, it is also a resource for *aligning* to the signals of other participants. Furthermore, we found that *early* alignment tends to be multimodal rather than unimodal. This may be because most referents were hard to describe, putting pressure on people to use both multimodal utterances and alignment as resources to establish joint reference early on in the interaction (which then yields early emergence of alignment in at least one modality). Conversely, for easier referents, both the need for alignment and multimodal communication could be lower (yielding later emergence of unimodal alignment).

Turning to the order of emergence for multimodal alignment, we found that simultaneous emergence of alignment in both modalities was most frequent, again underscoring the need to consider multimodal origins of language. However, we also found ample cases where alignment emerged in one modality first and later in the other, but contrary to our expectations, the two orders were equally frequent. We hypothesized to find ample "gestural alignment first", as this would resemble patterns in contexts of language development and language emergence where gestures (paired with vocalizations) can "pave the way" for the emergence of conventionalized lexical items (Fay et al., 2013; Iverson & Goldin-Meadow, 2005; Perry et al., 1988). While the quantitative finding that "lexical alignment first" was not rare was surprising, our qualitative analyses revealed that this occurred to deal with particular communicative challenges, as we will argue later in the discussion.

Qualitative findings

Our qualitative findings demonstrate *how* (multi)modality and alignment interact in collaborative referring. The results corroborate earlier work showing that alignment can be employed to accept or further negotiate a referring expression, which can be done through lexical alignment (Clark & Brennan, 1991; Clark & Wilkes-Gibbs, 1986), but also gestural alignment (Chui, 2014; Holler & Wilkin, 2011), or – as we showed here – by aligning in both modalities simultaneously. But our results bring to light another function as well: when various candidate expressions have been used for a referent, alignment can be used to commit to one of those conceptualizations.

A second insight from the qualitative analyses is that people employ both similarity *and* variation in gesture form for communicative purposes. We find evidence for what appears to be "strategic" alignment of communicatively "significant" form features (Bergmann & Kopp, 2012), where the sequential context governs which features (e.g., handshape, motion) are relevant at that moment. But our results also bring to light an alternative strategy: speakers can communicatively employ misalignment or deviation in salient form features to negotiate referring expressions (cf., Chui, 2014; Tabensky, 2001; see also Fusaroli et al., 2014; Healey et al., 2014 on this notion of complementarity in interaction). And finally, people might communicatively employ alignment of less significant form features as well, as a way to mark the common ground before adding new information. Transcript 1 provides an example of how these latter two strategies are combined: a participant repeated their partner's gesture with the same (non-salient) handedness, position, orientation and handshape (constituting the link to their partners gesture), but changed the movement into a salient, circular motion (to further specify the shape of the "plateau").

The analyses revealed that people employ modality-specific features when aligning. Whereas the discrete combinatorial format of speech allows for extending or modifying parts of noun phrases, the iconic and dynamic nature of gestures allows for copying or modifying form features to bring certain aspects of the referent in focus. These different affordances also enable people to balance the communicative load between the lexical and gestural modalities depending on the interactional needs at hand. Though overall the emergence of lexical alignment was more frequent, we also showed cases of how gestural alignment is used for achieving mutual understanding in the absence of a lexical pact (Transcript 7), or even in the absence of content words all together (Transcript 6). Gestural alignment also emerged when people experienced problems producing a verbal reference (where gestural alignment preceded lexical alignment; e.g., Transcript 2) or recalling an already established lexical pact (in which case gestural alignment follows lexical alignment; e.g., Transcript 4).

In summary, the spoken and gestural modalities offer their own affordances for alignment to establish joint reference, and these modalities are usually employed in combination. Our qualitative analyses help to make sense of the nuanced patterns that emerge from the quantitative findings. While the primacy of multimodal alignment emerges clearly throughout the study, the relative order of its building blocks, lexical and gestural alignment, appears to be governed by an interaction between the moment-by-moment communicative demands and the affordances offered by each modality.

Future research

Coming back to the initial question of how alignment and communicative modality are employed for establishing shared symbols, three challenges remain to be further explored: 1) how to operationalize alignment, 2) how to generalize the results, and 3) how to account for variation in shared symbol creation.

In order to systematically track both lexical and gestural alignment, we formulated maximally commensurate measures of what constitutes alignment, regarding behavior as aligned when it was produced in the same *modality* and for the same *referent*, and with modality-specific criteria for the required overlap in *form*. Our quantitative results should be interpreted and compared to prior work with this specific operationalization kept in mind. Specifically, while most studies on gestural alignment emphasize overlap in gesture form (Rasenberg et al., 2020), here we considered form overlap loosely. By pairing this with both a quantitative (see Appendix B) and qualitative investigation of gesture form overlap, we revealed how overlap *and* deviation in gesture form can be employed for communicative purposes. Future work could broaden the definition of alignment even further by also investigating cases where people verbally re-encode the information that their partner provided through gesture, or vice versa – that is, investigate alignment *across* modalities (de Fornel, 1992; Rasenberg et al., 2020; Tabensky, 2001).

As to the issue of generalizability, our dataset appears to be representative of this kind of task-based setting, as we find the same phenomena as described in earlier work using similar tasks (e.g., emergence of conceptual pacts, shorter references over time, vast amounts of iconic gestures; e.g., Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Holler & Wilkin, 2011). While the interactions are clearly different from everyday conversations, they do fulfill all basic characteristics of face-to-face conversation (Clark, 1996; see also the discussion by Holler & Wilkin, 2011) and show resemblances with common communicative situations, such as singling out a familiar referent from a set of similar referents (e.g., asking for a specific cup from a set of cups in a cupboard), or talking about novel objects or concepts (e.g., when working on an art project). Furthermore, since the Fribbles lack conventionalized labels, our data enabled us to shed some light on the potential interplay between alignment and modality in emergence contexts.

Lastly, we found quite some variation in the degree of shared symbol emergence, that is, the similarity of the names after the interaction. This variation could not be explained with the patterns of alignment in our data. This may be due to our focus on the *emergence* of alignment (i.e., the first occurrence), as opposed to repeated use (*entrainment*) later on in the interaction (see also Appendix C). Variation in systematicity and efficiency of novel symbols has previously been linked to the presence viz. absence of interactive feedback (Fay et al., 2018; Krauss & Weinheimer, 1966; Motamedi et al., 2019). Given that participants were allowed to interact as much as they wanted in our task, why did this not always give rise to simple, shared symbols as measured post-interaction? Future studies could explore this question further by investigating the kind of interactional work that is needed to go from the first occurrence of alignment to entrainment and simplification of shared symbols.

Conclusion

By systematically tracking lexical and gestural alignment in a referential communication task in a clearly operationalized way, we uncovered the primacy and prevalence of multimodal alignment when referring to novel objects. Moreover, by closely inspecting the interactional dynamics of independent, simultaneous, and successive emergence of lexical and gestural alignment, we found that the multimodal system can be flexibly adjusted to communicative pressures and constraints to yield referring expressions that contribute toward the ultimate goal of achieving joint reference. We believe a combination of qualitative and quantitative analyses akin to those in the present study have the potential to provide more insights into the joint contribution of different modalities (speech and gesture) in alignment of communicative behavior when creating novel symbols.

Notes

1. For the model comparing gestural alignment only to multimodal alignment, we only included a random intercept for subparts (not for dyads), due to convergence issues.
2. Though we did not have a hypothesis about the difference in frequency of multimodal alignment and lexical alignment only, we compared them to provide a complete picture and found no statistical difference (β = 0.70, *SE* = 0.74, *z* = 0.94, *p* = .348).
3. Note that the category gestural alignment only is rather small (*n* = 16); however, when comparing multimodal alignment to unimodal alignment (thus collapsing lexical alignment only and gestural alignment only), the distributions were significantly different as well (*p* = .002).
4. For the models comparing lexical alignment first to gestural alignment first and to simultaneous emergence, we only included a random intercept for dyads (not for subparts), due to convergence issues.
5. Though we did not have a hypothesis about the difference in frequency of simultaneous emergence and gestural alignment first, we compared them to provide a complete picture and found that simultaneous emergence was more frequent (β = 0.59, *SE* = 0.25, *z* = 2.33, *p* = .02).
6. Note that in our study we were not able to relate patterns of alignment to task performance, as all dyads scored at or near ceiling in the referential task.
7. 80% of all referentially aligned gestures overlap in at least one of the four features considered (handedness, handshape, movement, and orientation), but as many as 94% when also including position.

Acknowledgments

This paper came about in the context of a multidisciplinary research project within the Language in Interaction consortium, called Communicative Alignment in Brain and Behaviour (CABB). We wish to make explicit that the work has been shaped by contributions of CABB team members (in alphabetical order): Mark Blokpoel, Lotte Eijk, Judith Holler, Branka Milivojevic, Wim Pouw, Iris van Rooij, Ivan Toni and Marieke Woensdregt. We would additionally like to thank Wim Pouw and Marieke Woensdregt for their comments on an earlier version of this paper, Wim Pouw, Susanne Brouwer and the M3 members of the Radboud University for help with statistics, and Emma Berensen, Junfei Hu, Vicky Ioannidou, and Clarissa de Vries for their help with annotating the data. The authors remain solely responsible for the contents of the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant [024.001.006] to the Language in Interaction consortium and Vidi Grant [016.vidi.185.205] to M.D.

ORCID

Marlou Rasenberg http://orcid.org/0000-0003-1812-6907
Asli Özyürek http://orcid.org/0000-0002-0914-8381
Sara Bögels http://orcid.org/0000-0002-4945-5765
Mark Dingemanse http://orcid.org/0000-0002-3290-5723

Data availability statement

The data files and analysis scripts that support the findings of this study are openly available on the Donders Repository at https://doi.org/10.34973/7kbd-5g86.

References

Barry, T. J., Griffith, J. W., De Rossi, S., & Hermans, D. (2014). Meet the Fribbles: Novel stimuli for use within behavioural research. *Frontiers in Psychology*, *5*, 103. https://doi.org/10.3389/fpsyg.2014.00103

Bergmann, K., & Kopp, S. (2012). Gestural alignment in natural dialogue. In N. Miyake, D. Peebles, and R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* Sapporo Japan (pp. 1326–1331). Cognitive Science Society.

Bertrand, R., Ferré, G., & Guardiola, M. (2013). French face-to-face interaction: Repetition as a multimodal resource. In M. Rojc, and N. Campbell (Eds.), *Coverbal synchrony in human-machine interaction* (pp. 141–172). Science Publishers/CRC Press. https://hal.archives-ouvertes.fr/hal-00908190

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13–B25. https://doi.org/10.1016/S0010-0277(99)00081-5

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482

Brown, G., Anderson, A., Schillock, R., & Yule, G. (1984). *Teaching talk*. Cambridge University Press.

Cheney, D. L., & Seyfarth, R. M. (2005). Constraints and preadaptations in the earliest stages of language evolution. *The Linguistic Review*, *22*(2–4), 135–159. https://doi.org/10.1515/tlir.2005.22.2-4.135

Chui, K. (2014). Mimicked gestures and the joint construction of meaning in conversation. *Journal of Pragmatics*, *70*, 68–85. https://doi.org/10.1016/j.pragma.2014.06.005

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H. (2016). Depicting as a method of communication. *Psychological Review*, *123*(3), 324–347. https://doi.org/10.1037/rev0000026

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (Vol. 13, pp. 127–149). American Psychological Association.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, *49*(2), 214–230. https://doi.org/10.1016/S0749-596X(03)00060-3

Clift, R. (2016). *Conversation analysis*. Cambridge University Press.

Couper-Kuhlen, E., & Selting, M. (2017). *Interactional linguistics: Studying language in social interaction*. Cambridge University Press.

de Fornel, M. (1992). The return gesture: Some remarks on context, inference, and iconic gesture. In P. Auer, and A. Di Luzio (Eds.), *The contextualization of language* (pp. 159–176). John Benjamins Publishing Company. https://ci.nii.ac.jp/naid/10009705434/

Dideriksen, C., Christiansen, M. H., Tylén, K., Dingemanse, M., & Fusaroli, R. (2020). *Building common ground: Quantifying the interplay of mechanisms that promote understanding in conversations* (PsyArXiv). https://doi.org/10.31234/osf.io/a5r74

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G., & Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PLOS ONE*, *10*(9), e0136100. https://doi.org/10.1371/journal.pone.0136100

Duran, N. D., Paxton, A., & Fusaroli, R. (2019). ALIGN: Analyzing linguistic interactions with generalizable techNiques - A Python library. *Psychological Methods*, *24*(4), 419–438. https://doi.org/10.1037/met0000206

Fay, N., Arbib, M., & Garrod, S. (2013). How to bootstrap a human communication system. *Cognitive Science*, *37*(7), 1356–1367. https://doi.org/10.1111/cogs.12048

Fay, N., Lister, C. J., Ellison, T. M., & Goldin-Meadow, S. (2014). Creating a communication system from scratch: Gesture beats vocalization hands down. *Frontiers in Psychology*, *5*, 354. https://doi.org/10.3389/fpsyg.2014.00354

Fay, N., Walker, B., Swoboda, N., & Garrod, S. (2018). How to create shared symbols. *Cognitive Science*, *42*(S1), 241–269. https://doi.org/10.1111/cogs.12600

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological Science*, *23*(8), 931–939. https://doi.org/10.1177/0956797612436816

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147–157. https://doi.org/10.1016/j.newideapsych.2013.03.005

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, *40*(1), 145–171. https://doi.org/10.1111/cogs.12251

Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M. H., & Dingemanse, M. (2017). Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. In G. Gunzelmann, A. Howes, T. Tenbrink, and E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* London, UK (pp. 2055–2060). Cognitive Science Society.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218. https://doi.org/10.1016/0010-0277(87)90018-7

Goldin-Meadow, S. (2017). What the hands can tell us about language emergence. *Psychonomic Bulletin & Review*, *24*(1), 213–218. https://doi.org/10.3758/s13423-016-1074-x

Graziano, M., Kendon, A., & Cristilli, C. (2011). 'Parallel gesturing' in adult-child conversations. In G. Stam, and M. Ishino (Eds.), *Integrating gestures: The interdisciplinary nature of gesture* (pp. 89–101). John Benjamins Publishing Company.

Healey, P. G. T., Purver, M., & Howes, C. (2014). Divergence in dialogue. *PLOS ONE*, *9*(6), e98598. https://doi.org/10.1371/journal.pone.0098598

Holler, J., & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, *35*(2), 133–153. https://doi.org/10.1007/s10919-011-0105-6

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, *137*(2), 297–315. https://doi.org/10.1037/a0022128

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, *16* (5), 367–371. https://doi.org/10.1111/j.0956-7976.2005.01542.x

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–267. https://doi.org/10.1177/0956797609357327

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Kimbara, I. (2006). On gestural mimicry. *Gesture*, *6*(1), 39–61. https://doi.org/10.1075/gest.6.1.03kim

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32. https://doi.org/10.1016/S0749-596X(02)00505-3

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*(3), 343–346. https://doi.org/10.1037/h0023705

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130302. https://doi.org/10.1098/rstb.2013.0302

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior Matching in Multimodal Communication Is Synchronized. Cognitive Science, 36(8), 1404–1426. https://doi.org/10.1111/j.1551-6709.2012.01269.x

Lücking, A., Bergman, K., Hahn, F., Kopp, S., & Rieser, H. (2013). Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, *7* (1–2), 5–18. https://doi.org/10.1007/s12193-012-0106-8

Lücking, A., Ptock, S., & Bergmann, K. (2012). Assessing agreement on segmentations by means of staccato, the segmentation agreement calculator according to thomann. In E. Efthimiou, G. Kouroupetroglou, & S.-E. Fotinea (Eds.), *Gesture and sign language in human-computer interaction and embodied communication* (pp. 129–138). Springer Berlin Heidelberg.

MacNeilage, P. (2008). *The origin of speech*. OUP Oxford.

Macuch Silva, V., Holler, J., Ozyurek, A., & Roberts, S. G. (2020). Multimodality and the origin of a novel communication system in face-to-face interaction. *Royal Society Open Science*, *7*(1), 182056. https://doi.org/10.1098/rsos.182056

Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, *91*, 5–27. https://doi.org/10.1016/j.jml.2016.03.009

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

Mithen, S. J. (2005). *The singing neanderthals: The origins of music, language, mind and body*. Harvard University Press.

Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language*, *66*(1), 249–264. https://doi.org/10.1016/j.jml.2011.07.004

Mondada, L. (2018). Multiple temporalities of language and body in interaction: Challenges for transcribing multimodality. *Research on Language and Social Interaction*, *51*(1), 85–106. https://doi.org/10.1080/08351813.2018.1413878

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. https://doi.org/10.1016/j.cognition.2019.05.001

Oben, B., & Brône, G. (2016). Explaining interactive alignment: A multimodal and multifactorial account. *Journal of Pragmatics*, *104*, 32–51. https://doi.org/10.1016/j.pragma.2016.07.002

Özer, D., & Göksun, T. (2020). Gesture use and processing: A review on individual differences in cognitive resources. *Frontiers in Psychology*, *11*, 573555. https://doi.org/10.3389/fpsyg.2020.573555

Perlman, M. (2017). Debunking two myths against vocal origins of language: Language is iconic and multimodal to the core. *Interaction Studies*, *18*(3), 376–401. https://doi.org/10.1075/is.18.3.05per

Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: From a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130300. https://doi.org/10.1098/rstb.2013.0300

Perry, M., Breckinridge Church, R., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development*, *3*(4), 359–400. https://doi.org/10.1016/0885-2014(88)90021-4

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190. https://doi.org/10.1017/S0140525X04000056

Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in multimodal interaction: An integrative framework. *Cognitive Science*, *44*(11), e12911. https://doi.org/10.1111/cogs.12911

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29–46. https://doi.org/10.1016/j.jml.2014.05.008

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis* (Vol. 1). Cambridge University Press.

Sterelny, K. (2012). Language, gesture, skill: The co-evolutionary foundations of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2141–2151. https://doi.org/10.1098/rstb.2012.0116

Tabensky, A. (2001). Gesture and speech rephrasings in conversation. *Gesture*, *1*(2), 213–235. https://doi.org/10.1075/gest.1.2.07tab

Zlatev, J., Wacewicz, S., Żywiczyński, P., & van de Weijer, J. (2017). Multimodal-first or pantomime-first?: Communicating events through pantomime with and without vocalization. *Interaction Studies*, *18*(3), 465–488. https://doi.org/10.1075/is.18.3.08zla

Appendix A Inter-rater reliability for gesture coding

To establish inter-rater reliability for gesture coding, we focused on the first two rounds of the interaction (where presumably the most (diverse) gestures would occur), where two coders independently coded 15% of the trials (96 trials, n = 296 gestures). Inter-rater agreement on gesture identification was 89.2%. For this measure, we scored how many annotations overlapped, where we disregarded differences in handedness, the length of the annotations, and/or the number of segments (e.g., one stroke annotation from one coder spanning two stroke annotations of the other coder). To also assess these aspects of the degree of organization of the coder's segmentations, we used the Staccato algorithm (Lücking et al., 2013, 2012). We applied this to the left and right hand of each speaker separately, which resulted in a mean score of 0.76 (on a scale from −1 to 1) – indicating that the coders had similar understandings of how the observed gestures had to be segmented. Inter-rater agreement for gesture type was substantial (agreement = 95.1%, Cohen's kappa = .64) and for gesture referent high (agreement = 92.8%, Cohen's kappa = .93).

Appendix B Operationalization of alignment: form similarity

Lexical alignment

Lexical alignment is coded per Fribble subpart, using the same referent coding procedure as described for gestures (see Figure 3 in manuscript). As for the form criterium: we consider words to be aligned if they have the same root form (or "lemma"), so diminutive or plural forms count as aligned, but synonyms or paraphrases do not (cf., Oben & Brône, 2016). Participants sometimes align on multiple words (e.g., both refer to a subpart with "flat nose"), but lexical alignment is computed as a binary variable where alignment of *one* lemma suffices. The specific categories of words that are included and excluded are listed in Table B1.

Table B1. Categories of lexical items included and excluded in the analysis of lexical alignment.

Category	Examples (English translations)
Included	
Shape	*circle, cone, hook, trunk, round, elongated*
Size	*small, big, mini*
Orientation	*upright, diagonal, downwards*
Manner of attachment	*against, through, sticking out, surrounding it*
Similarities/differences between subparts	*two, three, the same, different*
Excluded	
Non-referential speech	meta-speech about the task, such as *"oh we're getting better at this!"*
Highly frequent words*	verbs *to have* and *to be*, as well as most pronouns, determiners and conjunctions
Hedging	*sort of, kind of, little bit, like*
Non-informative speech that applies to all Fribbles	words related to general positions, such as *left, right, on top of;* as well as generic words to describe subparts such as *shape, figure, thing*.

* Frequency was determined on the basis of the SUBTLEX-NL corpus (Keuleers et al., 2010), where we used three standard deviations from the mean lemma frequency as the cutoff for "high frequency".

Gestural alignment

Prior work has used various *form* criteria for considering gestures as aligned. For instance, gestures should have the same representation technique (e.g., drawing or handling; Oben & Brône, 2016) and/or the same "overall form" (Holler & Wilkin, 2011; Bertrand et al., 2013 for a review, see Rasenberg et al., 2020). Based on an explorative analysis on overlap in gesture form features in our data (as described in the next paragraph), we have decided to include *all* referentially aligned gestures, irrespective of the degree of form similarity. The reasoning, in a nutshell, is that for a well-motivated set of basic form features, a great majority of candidate aligned gestures in our data showed similarities on one or more features, making the set of all candidate gestures a reasonable proxy for form-aligned gestures.

For a subset of the referentially aligned gesture pairs (for 8 dyads in round 1 and round 2, n = 389 gestures), gestures were coded for their similarity in form. This study is the first to provide such a quantitative analysis of form similarity for a relatively large set of gestures which are related in meaning (see Chui, 2014 for a similar approach with a small sample; Bergmann & Kopp, 2012 for a large-scale quantification of form similarity for gestures based on their temporal rather than semantic relation). Similarity was coded in terms of five form features: handedness, handshape, movement, orientation, and position in a binary fashion. Coding was done by a trained assistant who was naive to the study's rationale. That is, the coder saw the gesture stroke annotations along with the videos, but had no access to the co-occurring speech or referent coding, and was blind to the selection procedure of the gesture pairs.

Inter-rater reliability for gesture form similarity coding was assessed separately for the five features. Agreement for handedness was computed based on 15% of the initial gesture annotations in rounds 1 and 2 (see the section *gesture type and referent*), and resulted in high agreement (agreement = 94.7%, Cohen's kappa = .91). For the other features, a second trained, naive assistant coded 25% of the referentially aligned gesture pairs (n = 103) for overlap in handshape, movement, orientation, and position. Substantial agreement was obtained for handshape (agreement = 88.3%, Cohen's kappa = .71) and movement (agreement = 85.4%, Cohen's kappa = .63), and moderate to substantial agreement for orientation (agreement = 75.7%, Cohen's kappa = .54). For position, the score was on the lower side of the moderate category (agreement = 77.7%, Cohen's kappa = .47), and so this feature was excluded from further analyses.

The results of the explorative analyses are shown below. As becomes apparent from Figure B1 (panel A), overlap in handedness appears to be most frequent (which naturally follows from the limited degree of freedom: gestures are either left-handed, right-handed or two-handed). Panel B shows that for the number of features that overlap in each gesture pair, overlap in two features is most common and "complete" form overlap (similar on all four form features) is rare for most dyads. Overall, 80% of all gesture pairs have partial form overlap (similar on one or more features), while only 4% has complete form overlap. Figure B2 shows a combination of the plots in Figure B1: it shows *which* features are most likely to overlap for gesture pairs with a particular number of overlapping features (1, 2 or 3).

In conclusion, based on the fact that the majority of gesture pairs show at least *partial* form overlap,[7] we included all referentially aligned gesture pairs irrespective of their form features.

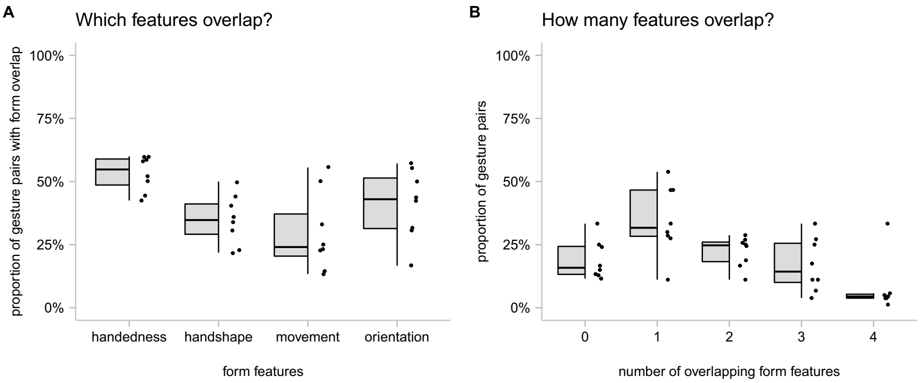


Figure B1. Form similarity of referentially-aligned gestures. Panel **A** shows the relative frequencies with which each form feature overlaps. Panel **B** shows the relative frequencies of the number of features that overlap. Dots represent dyads (N = 10).

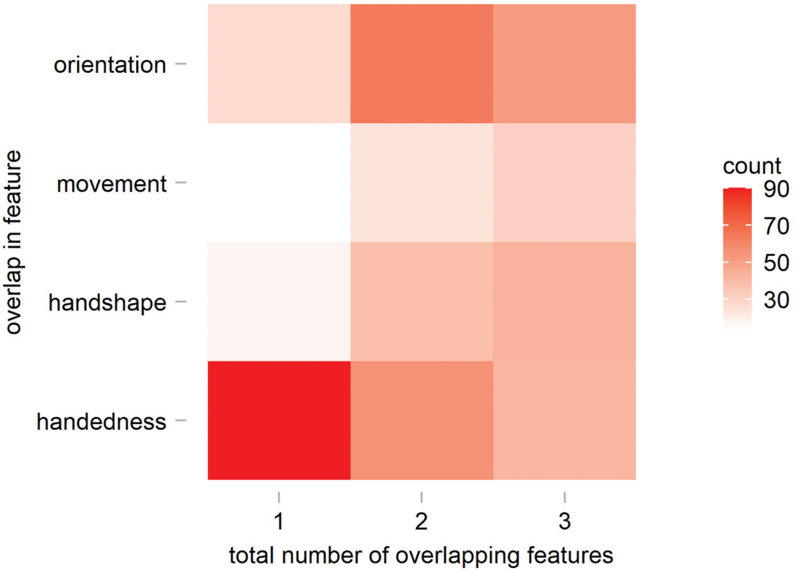


Figure B2. Heatmap displaying *which* features are most likely to overlap for gesture pairs with a particular number of overlapping features (1, 2, or 3).

Appendix C Relation between alignment and naming similarity scores

In the main article we state that there is no evident relationship between the degree of lexical and gestural alignment in the interaction and the post naming similarity scores. Here, we present a more detailed inspection of these variables, and the relation between them.

Post-naming similarity

Both before and after the interaction, participants were asked to individually label the Fribbles (target objects) such that their partner could find them. Figure C1 shows the distribution of the naming similarity scores *post* interaction. Though oftentimes dyads provide similar or even exactly the same names, there is also a large amount of naming pairs that have zero similarity after the interaction (we elaborate on this in section "**Shared symbols in the naming task"** in the manuscript).

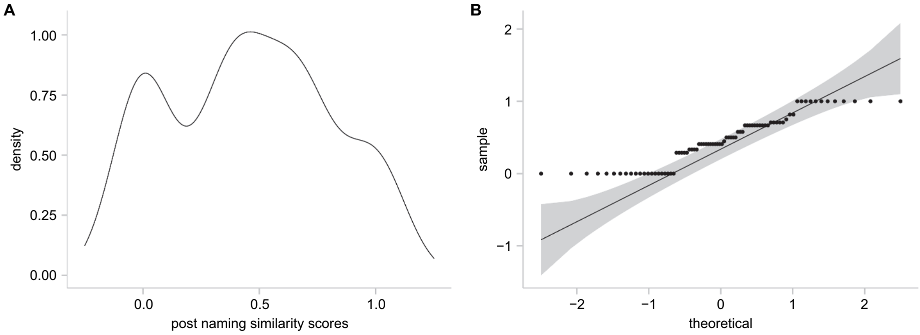


Figure C1. Density plot (panel **A**) and quantile-quantile plot (panel **B**) for *post* naming similarity scores (N = 80).

Alignment

First, note that rather than counting the overall frequencies of alignment over the whole interaction, we have specifically tracked the first occurrence of alignment in each modality. That is, we have quantified how often alignment *emerged* in the lexical and/or gestural modality for particular referents (and did not track repeated usage later in the interaction). Second, while naming scores were computed per Fribble (on a scale from 0 to 1), alignment was measured (categorically) for Fribble *subparts*. To be able to relate these variables, we took the relative number of Fribble subparts per Fribble for which alignment emerged as the "degree of alignment" per Fribble. We summed all categories of alignment here (lexical only, gestural only, and multimodal). Including them separately would have resulted in multicollinearity, because the (mutually exclusive) categories are not independent of each other. For example, if for a particular dyad all subparts of a Fribble were grouped in the category multimodal alignment, then it naturally follows that there were zero subparts in the category lexical alignment only.

As shown in Figure C2, alignment tended to emerge for almost all subparts (*Median* = 100%).

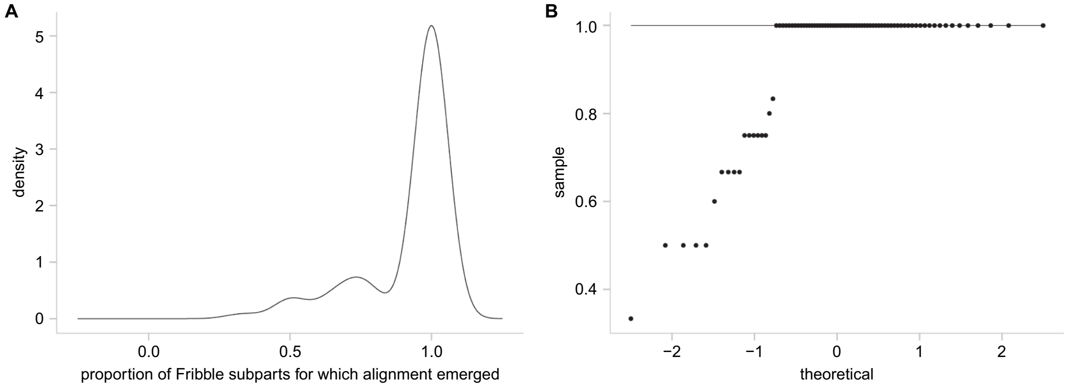


Figure C2. Density plot (panel **A**) and quantile-quantile plot (panel **B**) for the proportion of Fribble subparts for which alignment emerged (*N* = 80).

Relation between alignment and post-naming

Figure C3 displays the relation between the relative number of Fribble subparts for which alignment occurred and the *post* naming similarity scores, and shows that there is no evident relationship between the two. The fact that we only measured *emergence* of alignment might explain why we do not find a relation with *post* naming scores, but as becomes clear from Figure C3 this is further complicated by the fact that alignment scores (as measured per Fribble) are near ceiling.

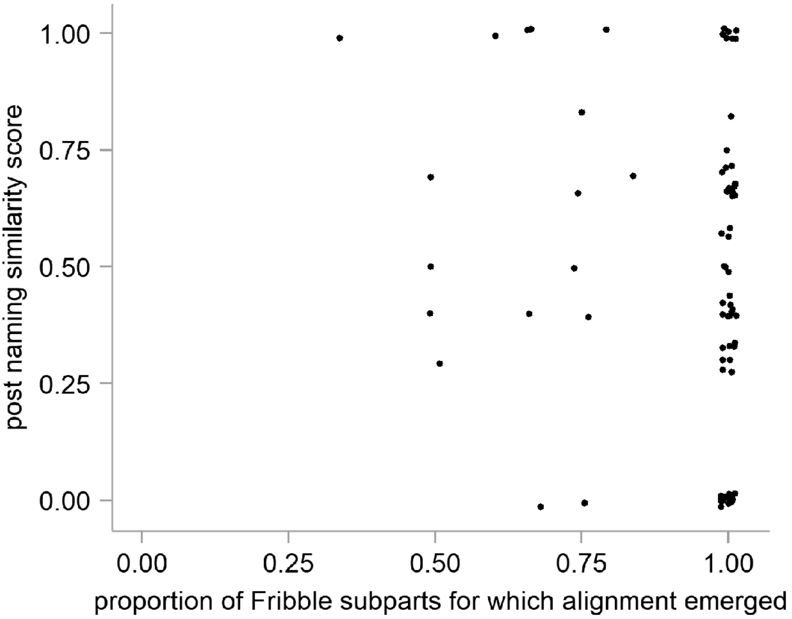


Figure C3. Scatterplot showing the relation between the relative number of Fribble subparts for which alignment emerged and the *post* naming similarity scores (*N* = 80).