Speakers align both their gestures and words not only to establish but also to maintain reference to create shared labels for novel objects in interaction

Sho Akamine (sho.akamine@mpi.nl)
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Esam Ghaleb (e.ghaleb@uva.nl)
Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands

Marlou Rasenberg (marlou.rasenberg@meertens.knaw.nl)
Meertens Institute, Amsterdam, The Netherlands

Raquel Fernández (raquel.fernandez@uva.nl)
Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands

Antje Meyer (antje.meyer@mpi.nl)
Aslı Özyürek (asli.ozyurek@mpi.nl)
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
Donders Centre for Brain, Cognition, and Behavior, Radboud University, Nijmegen, The Netherlands

Abstract

When we communicate with others, we often repeat aspects of each other's communicative behavior such as sentence structures and words. Such behavioral alignment has been mostly studied for speech or text. Yet, language use is mostly multimodal, flexibly using speech and gestures to convey messages. Here, we explore the use of alignment in speech (words) and co-speech gestures (iconic gestures) in a referential communication task aimed at finding labels for novel objects in interaction. In particular, we investigate how people flexibly use lexical and gestural alignment to create shared labels for novel objects and whether alignment in speech and gesture are related over time. The present study shows that interlocutors establish shared labels multimodally, and alignment in words and iconic gestures are used throughout the interaction. We also show that the amount of lexical alignment positively associates with the amount of gestural alignment over time, suggesting a close relationship between alignment in the vocal and manual modalities.

Keywords: behavioral alignment; multimodal language; iconic gestures; referential communication

Introduction

In daily conversation, people repeat aspects of each other's communicative behavior, such as sentence structures, words, eye gaze, and gestures (Pickering & Garrod, 2004; Rasenberg et al., 2020). This cross-participant repetition of communicative behavior is called *behavioral alignment*.

Behavioral alignment has been shown to occur at different linguistic levels. For example, Levelt and Kelter (1982) showed that participants are more likely to answer a question with a prepositional phrase (e.g., "*At five o'clock*") when the question contains a preposition ("*At what time does your shop close?*") compared to when it doesn't ("*What time does your shop close?*"). On the other hand, when the question did not contain a preposition, they tended to produce an answer without a preposition. Further, question-answer pairs were judged as more natural when they both contained or did not contain prepositions than when one contained a preposition while the other did not. This suggests that people prefer to reuse the same lexicosyntactic structures as preceding utterances. Alignment in syntactic structures has also been demonstrated in other studies (e.g., Hartsuiker et al., 2008; Ivanova et al., 2020), although when adjusted for word repetition, syntactic alignment did not exceed chance levels in corpus-based studies (Green & Sun, 2021; Healey et al., 2014).

Not only does alignment occur in syntactic structures but also in lexical items. A salient example of lexical alignment is that when one speaker refers to a type of seat using the word "*sofa*", the next speaker is more likely to refer to it as "*sofa*" instead of "*couch*" (Branigan et al., 2010).

The interactive alignment model (Pickering & Garrod, 2004) suggests that alignment at one linguistic representation level can influence and lead to alignment at other levels, which can be reflected in the interlocutor's utterances as behavioral alignment. This idea has been supported for syntactic, lexical, and semantic alignment (Mahowald et al., 2016). While the original model does not incorporate the visual modality (e.g., facial expressions, manual gestures), the authors expanded it to incorporate co-speech gestures as a level of linguistic representation in their recent publication (Pickering & Garrod, 2021). However, the relationships between alignment in the vocal and visual modality remain unexplored. Here, we ask whether and to what extent alignment in speech (words) and co-speech gestures go hand-in-hand and related over time when interlocutors are referring to novel objects in a referential communication task.

Previous studies suggest that alignment plays a functional role in reducing the processing cost associated with language production (Bartolozzi et al., 2021; Norrick, 1987), signaling agreement or disagreement (Norrick, 1987; Põldvere et al., 2021), signaling (lack of) understanding (Crible et al., 2024; Norrick, 1987), and establishing a partner-specific temporal agreement on how to conceptualize novel objects (i.e., *con-*

ceptual pacts), to name a few (Brennan & Clark, 1996; Clark & Brennan, 1991; Clark & Wilkes-Gibbs, 1986).

Previous alignment research has focused on alignment in speech. However, language involves not only vocal signals but also bodily signals such as manual gestures (e.g., Kendon, 2004; Kita & Özyürek, 2003; McNeill, 2005; Özyürek, 2018; Özyürek & Woll, 2019). Due to the multimodal nature of language, recent studies have explored alignment in co-speech gestures (Bergmann & Kopp, 2012; Holler & Wilkin, 2011; Kimbara, 2006, 2008; Oben & Brône, 2016). These studies have shown alignment can occur not only in speech but also in gestures. However, it is still unclear whether they go hand in hand and whether they stay over the course of interaction.

In a recent study using a referential communication paradigm for novel objects (Fribbles) in which participants communicated about each Fribble in six rounds, Rasenberg et al. (2022) showed that the emergence of first aligned expressions in the interaction is mostly multimodal (lexical + gestural alignment). When the first alignment between participants emerged, 56% involved both lexical and gestural alignment, while 38% and 6% involved lexical or gestural alignment, respectively. They also found that alignment mostly emerged in the initial two rounds. In the present study, we extend on Rasenberg et al. (2022) and investigate whether speakers maintain their alignments in both modalities throughout the interaction or whether gestural alignment served mainly to establish lexical alignment after which gestural alignment might not be needed.

The Present Study

Previous studies have demonstrated that interlocutors align in aspects of each other's speech (e.g., Dideriksen et al., 2022; Pickering & Garrod, 2004) and gestures (e.g., Holler & Wilkin, 2011; Oben & Brône, 2016; Rasenberg et al., 2022), and alignment mostly emerges as a multimodal phenomenon (Rasenberg et al., 2022). However, little is known about alignment in the vocal and manual modalities over the course of the interaction.

To further our understanding of people's use of alignment in speech and co-speech manual gestures when they communicate about novel referents, the present study investigates (1) whether interlocutors keep aligning in words and gesture after the emergence of lexical and gestural alignment, and (2) how alignment in speech and co-speech iconic gestures are related over time.

To investigate the first question, we examined the patterns of lexical and gestural alignment rates over the rounds of a referential communication task. Suppose the interlocutors maintain both lexical and gestural alignment after establishing the first instance of alignment. In that case, we expect the alignment rate to increase initially and stabilize for both lexical and gestural alignment. In contrast, if the interlocutors align in gestures mainly to establish lexical alignment but then maintain lexical alignment only because gestural alignment is no longer needed, we expect the alignment rate to increase initially and remain for lexical alignment throughout the interaction but decrease for gestural alignment.

As for the second research question about the relationship between alignment in vocal and manual modalities, we expect a positive association between lexical and gestural alignment rates if lexical and gestural alignment coincide. We hypothesize that lexical and gestural alignment typically coincide, as the interactive alignment model proposes that alignment at one linguistic representation level could influence and lead to alignment at another level, and linguistic representation levels include co-speech gestures (Pickering & Garrod, 2021).

Method

Dataset

The current study is based on a dataset collected in the context of the Communicative Alignment in Brain and Behaviour (CABB) research project. The dataset used here comprises a) a dyadic referential communication task with novel objects and b) a naming task, which participants individually performed before and after the communication task. The dataset was originally introduced in Rasenberg et al. (2022), in which they investigated the emergence of multimodal alignment in the referential communication task, analyzing a subset of data (half of the dyads and half of the stimuli items). The current study tracks lexical and gestural alignment as it develops over time and uses the entire dataset, that is, all the dyads and all the items. For a comprehensive description of the dataset and the speech and gesture annotations, see Rasenberg (2023).

Participants We analyzed data from 38 Dutch-speaking participants (20 women and 18 men, M_{age} = 22.5 years, $Range_{age}$ = 18–32 years). They were randomly grouped into 9 same-gender dyads and 10 mixed-gender dyads, each consisting of two unacquainted participants.

Materials The referential communication and naming tasks used a set of 16 pictures of blue 3D objects composed of a main body and 3–6 subparts (Figure 1). These objects, called Fribbles, were generated by modifying the original Fribbles created by Barry et al. (2014) to increase variation in elicited names (Eijk et al., 2022; Rasenberg et al., 2022).

Procedure Participants performed a referential communication task. The referential communication task consisted of 6 rounds, each with 16 trials. In each trial, participants were assigned the role of director or matcher. The director's role was to describe the target Fribble indicated by a red square around it, and the matcher's role was to identify it through free interaction. They were instructed that they were free to communicate however they wanted. Fribbles were presented on a 24-inch screen for each participant, with their order varying across participants. To avoid confusion about different Fribble orders, Fribbles were labeled with numbers for one participant and with letters for the other. Once the matcher was confident which Fribble the director was describing, they said the label for the Fribble and pressed a button to proceed

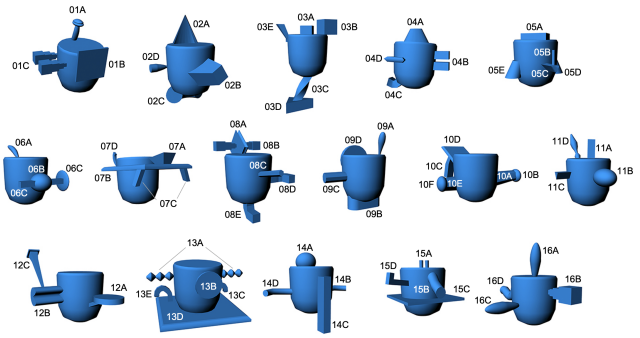

Figure 1: Sixteen Fribbles used as experimental stimuli for the referential communication task. Each subpart is labeled with Fribble numbers and alphabets. Note that participants did not see the labels for the subparts. They were labeled for coding purposes.

to the next trial. Once they completed all 16 trials, the order of Fribbles was shuffled, and the next round started. The interaction lasted for 24.46 minutes on average (range = 14.19–34.56 minutes). As the rounds proceeded, the trial duration became shorter (mean duration of rounds in minutes: R1 = 8.20, R2 = 5.00, R3 = 3.37, R4 = 3.00, R5 = 2.49, R6 = 2.40).

Speech Transcription and Gesture Annotation Speech transcription and gesture segmentation and coding were manually performed in ELAN (version 5.8; Wittenburg et al., 2006). The speech was segmented into Turn Construction Units (TCU; Couper-Kuhlen & Selting, 2018) and transcribed using conventional Dutch spellings. For co-speech gestures, only the stroke phase was annotated for each hand. The annotated gestures were manually categorized into three gesture types: iconic, deictic/pointing, and other gestures (e.g., beat gesture, interactive gesture). The annotation contained 71695 (28152 content) words and 4843 (4413 iconic) gestures.

Fribble subparts were used as a reference for iconic gestures. For instance, if an iconic gesture referred to the "antenna" of Fribble 1, the reference was coded as "01A". The reference coding was based on the iconic gesture, co-occurring speech, and context. For more details and inter-rater reliability measures for gesture annotation and coding, see Rasenberg (2023) and Rasenberg et al. (2022).

Analysis

Lexical Alignment Coding We defined lexical alignment as a cross-participant repetition of single or sequences of words containing at least one content word referring to the same target Fribble. The rationale for restricting our analyses to content words is that they tend to be more meaningful and unique to each Fribble (subpart) than function words (Brennan & Clark, 1996; Hawkins et al., 2020). As for the timing of alignment, we did not pose any restrictions: two expressions that refer to a particular Fribble were considered to be aligned regardless of the rounds in which they were produced as long as the first expression was produced by one of the interlocutors and the second by the other. Figure 2 shows a visual representation of lexical alignment operationalization.

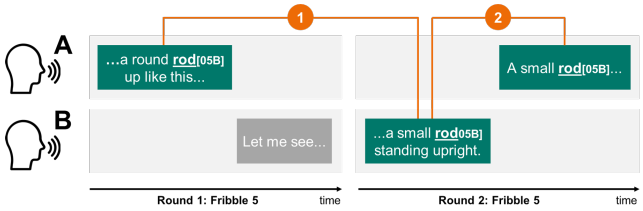

Figure 2: A visual representation of lexical alignment operationalization. Here, the word "*rod*" is a shared expression, and the numbers in orange circles represent lexical alignment counts. The numbers and letters in the square brackets represent Fribbles and their subparts. Note that the labels for the Fribble subpart were added for illustration purposes.

To prepare data for lexical alignment detection, we preprocessed speech transcripts in Python to remove disfluencies, mark part-of-speech (PoS), and lemmatize[1] each word (e.g., eating → eat). For PoS tagging and lemmatization, we used the spaCy library (Honnibal et al., 2020). Next, following a similar approach to Sinclair and Fernández (2021) and Ghaleb et al. (2024), we extracted all *shared expressions*: lemmatized expressions that both interlocutors used to describe a particular Fribble. For example, if both speakers A and B use the word "Pinocchio" to refer to Fribble 9, we regard "Pinocchio" as a shared expression for Fribble 9. Similarly, if it was used for multiple Fribbles (e.g., Fribble 8 and 9) by *both* speakers, we regard "Pinocchio" as a shared expression for Fribble 8 and 9. However, if speaker A uses "Pinocchio" only for Fribble 9, and speaker B uses the same word only for Fribble 8, then this is not a shared expression, as both speakers did not use the expression for the same Fribble. The rationale for this was to keep the alignment coding consistent for lexical and gestural alignment. Finally, we marked two shared expressions as an instance of lexical alignment if the first shared expression was produced by one of the interlocutors and the next one by the other interlocutor. The final dataframe contained 3351 instances of lexical alignment. Each row consisted of key attributes of the first and second components of the lexically aligned shared expressions pair.

Gestural Alignment Coding Gestural alignment was operationalized as a cross-participant repetition of iconic gestures referring to the same referent (i.e., Fribble subpart Rasenberg et al., 2022). Rasenberg et al. (2022) found that when two gestures used by both speakers referred to the same subpart, 80% of a subset of the aligned gesture pairs shared one or

[1]We are aware of the possibility that by lemmatizing words, we may miss instances of non-alignment. For instance, if we compare two cases where A says "eating" and B "eats" and where A and B both say "eats", the latter is more aligned because the form of lexical items is identical, while in the former case, the form of lexical items is different. However, we would argue that this concerns (non-)alignment in morphosyntactic structures rather than lexical items and does not affect our analyses of lexical alignment.

more form features out of four (i.e., handedness, handshape, movement, and orientation). As such, we posed no restriction on the gesture form to detect gestural alignment. Also, similar to lexical alignment, we did not pose any restrictions on the timing for gestural alignment. Figure 3 shows a visual representation of gestural alignment operationalization.

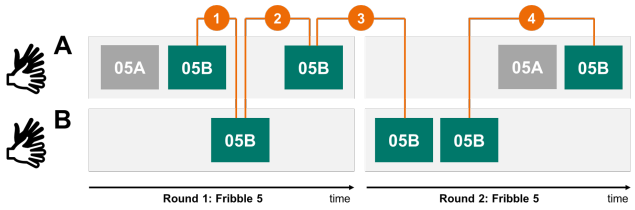

Figure 3: A visual representation of gestural alignment operationalization. The numbers and letters in the rectangles represent references to Fribbles and their subparts, respectively. The numbers in orange circles represent gestural alignment counts.

To detect gestural alignment, we first extracted iconic gestures and their key attributes, such as gesture referent, hand used, and timestamps. We then iterated through the data for each Fribble subpart, extracting all iconic gestures produced for the particular subpart. For the two-handed gestures referring to two subparts simultaneously (e.g., 14B+14D), we separated the referent into two independent subparts (e.g., 14B and 14D) before extracting iconic gestures for each referent. Lastly, we marked two iconic gestures produced for the same subpart as gestural alignment if the first one was produced by one speaker (e.g., speaker A) and the second by the other (e.g., speaker B). The final dataframe contained 1086 instances of gestural alignment. Each row consisted of key attributes of the first and second gestures of the aligned iconic gestures pair. Figure 4 shows an example of multimodal (lexical + gestural) alignment.

Statistical Analysis To analyze the data, we prepared separate dataframes for gestural and lexical alignment in which each row contains key attributes, including dyad number, trial number, trial duration, and the number of instances of alignment. Each dataset consisted of 1824 rows (19 pairs * 96 trials).

We performed zero-inflated multilevel Poisson regression models on the lexical and gestural alignment rate using the glmmTMB package (Brooks et al., 2017) in R (R Core Team, 2020). We use Poisson regression models, as "[t]he Poisson distribution is the canonical distribution for characterising count data with no or unknown upper bound" (Winter & Bürkner, 2021, p.1). Zero-inflated models assume that the "data contain more zeros than expected under the Poisson or negative binomial distributions, and these additional zeros are generated by separate processes (Winter & Bürkner, 2021)." Because there were trials in which participants produced no gestural or lexical alignment (1272 and 407 out of 1824, respectively), zero-inflated models were suitable for the data.

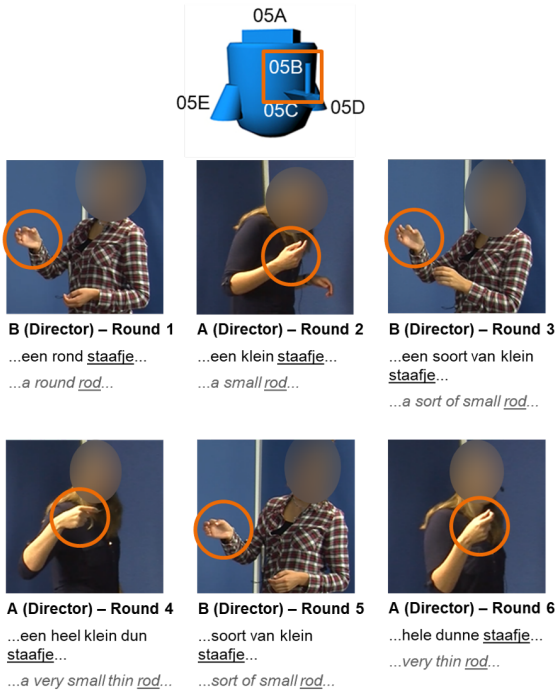

Figure 4: An example of multimodal (lexical + gestural) alignment over the six rounds. Here, both interlocutors used the word "*rod*" with an iconic gesture with vertical movements depicting the shape of Fribble subpart 05B.

The models included fixed effects for the round where alignment was established and random intercepts for dyads and target Fribbles, with the log-transformed trial duration as an offset variable (Winter & Bürkner, 2021). The offset variable was included to account for the longer trial duration in the initial rounds than in the later rounds, resulting in more opportunities for alignment in longer trials. It is worth mentioning that including the offset variable changes the unit of the response variable: the model with an offset variable for trial duration in minutes models the average amount of lexical or gestural alignment *per minute* (alignment rate). The formula for the multilevel Poisson regression models is as follows:

$$alignment\ count \sim round + (1|dyad) + (1|target) + \text{offset}(\log(trial\ duration\ min))$$

Results

Lexical Alignment Rate

A zero-inflated multilevel Poisson regression model showed a significant increase in the lexical alignment rate from round 1 to round 2 ($\beta = 0.99, SE = 0.06, z = 15.33, p < .01$) and round 2 to 3 ($\beta = 0.26, SE = 0.06, z = 4.54, p < .01$). We did not observe any significant increase or decrease in the later rounds (R3–R4: $\beta = 0.10, SE = 0.59, z = 1.76, p = .08$; R4–R5: $\beta = 0.09, SE = 0.06, z = 1.52, p = .13$; R5–R6: $\beta =$

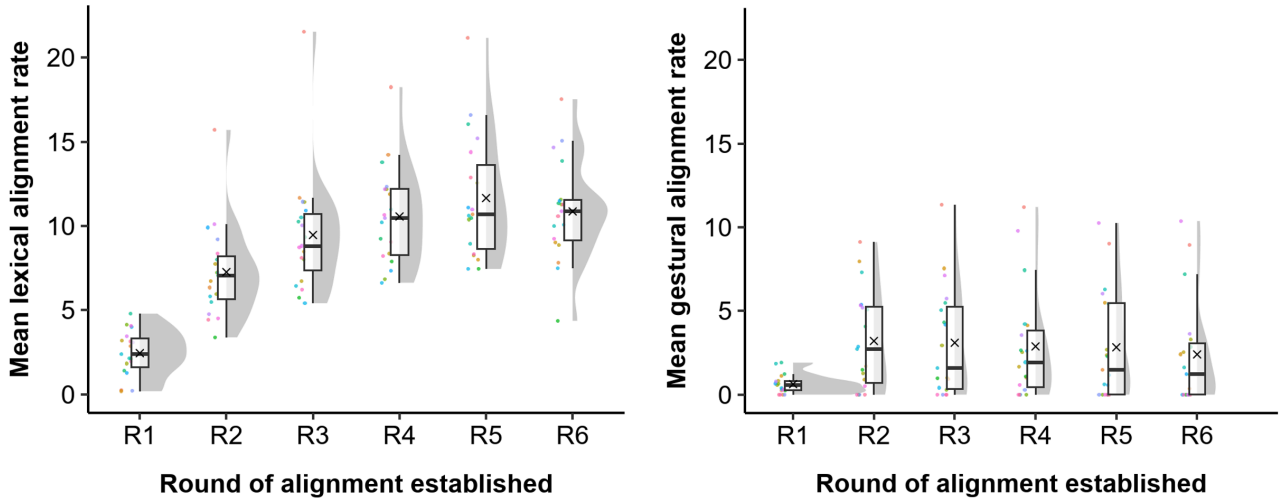

Figure 5: The distribution of alignment rate across rounds. The x-axis represents the round in whichvthe second component of aligning lexical items or gestures occurred, and the y-axis represents the alignment rate. The left figure illustrates lexical alignment rate, and the right figure illustrates gestural alignment rate. Each colored dot represents a dyad in a particular round. The Xs in the plot indicate the mean, and the midline in the boxplot represents the median.

$-0.07, SE = 0.06, z = -1.17, p = .24$). This suggests that the lexical alignment rate increases until round 3 and then stays in the later rounds. The left panel in Figure 5 shows the distribution of the lexical alignment rate across rounds.

Although the model tells us about the changes in the lexical alignment rate between two immediate rounds, it does not show whether it increased or decreased across multiple rounds. As we are interested in investigating whether the lexical alignment rate stayed or changed over the interaction, we changed the contrast coding of the round variable to a treatment coding with round 6 as the reference level and ran another zero-inflated multilevel Poisson regression model. The model revealed that compared to round 6, the number of alignment per minute was significantly lower in round 1 ($\beta = -1.39, SE = 0.07, z = -19.32, p < .01$) and 2 ($\beta = -0.38, SE = 0.06, z = -6.18, p < .01$). The model did not show significant differences between rounds 6 and 3 ($\beta = -0.13, SE = 0.06, z = -1.97, p = .05$[2]), rounds 6 and 4 ($\beta = -0.01, SE = 0.06, z = -0.23, p = .82$), and rounds 6 and 5 ($\beta = 0.07, SE = 0.06, z = 1.09, p = .27$).

In summary, the amount of lexical alignment increased in the initial rounds and stabilized afterward.

Gestural Alignment Rate

A zero-inflated multilevel Poisson regression model showed a significant increase in the number of gestural alignment per minute from round 1 to round 2 ($\beta = 1.41, SE = 0.11, z = 12.38, p < .01$). We did not observe any significant increase or decrease in the later rounds (R2–R3: $\beta = 0.01, SE = 0.09, z = 0.07, p = .94$; R3–R4: $\beta = -0.04, SE = 0.10, z = -0.40, p = .69$; R4–R5: $\beta = -0.05, SE = 0.11, z = -0.44, p = .66$; R5–R6: $\beta = -0.12, SE = 0.12, z = -0.96, p = .33$). This suggests that the amount of gestural alignment increases until round 2 and then remains stable in the later rounds (see the right panel in Figure 5 for the distribution of gestural alignment count per minute across rounds).

We also run another zero-inflated multilevel Poisson regression model with round 6 as the reference level to test if the number of gestural alignment stays or changes over the interaction. The model showed that compared to round 6, the number of gestural alignment per minute was significantly lower in round 1 ($\beta = -1.18, SE = 0.14, z = -8.37, p < .01$). However, we did not observe any other significant differences (R6–R2: $\beta = 0.15, SE = 0.12, z = 1.27, p = .21$; R6–R3: $\beta = 0.20, SE = 0.12, z = 1.60, p = .11$; R6–R4: $\beta = 0.21, SE = 0.13, z = 1.65, p = .10$; R6–R5: $\beta = 0.12, SE = 0.13, z = 0.87, p = .38$).

In summary, similar to lexical alignment, the amount of gestural alignment increased in the initial rounds and stabilized afterward. Although the general pattern was similar for lexical and gestural alignment, they differed as to when the alignment count stabilized. Taken together, the results are consistent with the prediction that interlocutors keep aligning in words and gestures throughout the interaction.

Association between Lexical and Gestural Alignment Rate

To test whether lexical alignment rate positively associates with gestural alignment rate, we performed a zero-inflated multilevel Poisson regression model with the response variable for lexical alignment rate, fixed effects for gestural alignment rate, random intercepts for dyad and target Fribble, and

[2]The alpha level is adjusted and set to 0.025 to control for the increased risk of the Type I error rate when performing multiple tests or models to test a single hypothesis (Chen et al., 2017; Rubin, 2021).

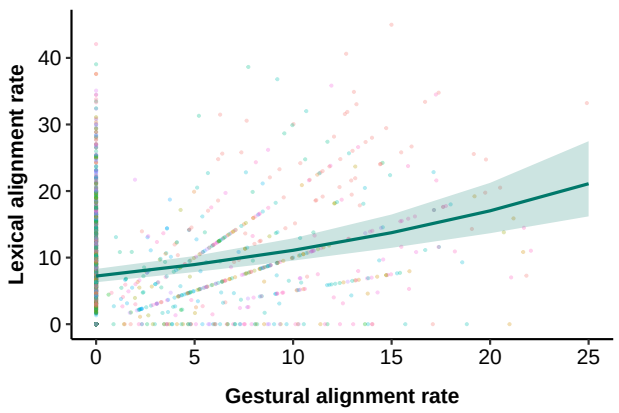

Figure 6: Lexical alignment rate as a function of gestural alignment rate. The colored dots represent each dyad's lexical and gestural alignment rate for each trial, and the green line indicates a predicted value based on the negative binomial regression model. The ribbon indicates the 95% confidence interval.

an offset variable for trial duration in minute. However, as we detected more variation in data than the model expected (i.e., *overdispersion*), we performed a zero-inflated multilevel negative binomial regression model with the same model structure (Winter & Bürkner, 2021).

A negative binomial regression model on the same dataframe we used in the previous sections, in which one data point corresponds to the amount of lexical and gestural alignment per trial (N = 1824), revealed that a higher gestural alignment rate is associated with a higher lexical alignment rate ($\beta = 0.04, SE = 0.01, z = 7.94, p < .01$; see Figure 6[3]). This is consistent with the hypothesis that lexical alignment tends to coincide with gestural alignment.

Discussion

In the present study, we investigated two questions: (1) whether interlocutors maintain their alignment in both the vocal and manual modalities and (2) how alignment in speech and gestures are related over time. As a step towards answering the questions, we examined the amount of lexical and gestural alignment produced in a copresent referential communication task.

RQ1: Do interlocutors maintain their alignment in both the vocal and manual modalities?

Statistical analyses on lexical and gestural alignment rates demonstrated that the alignment rate increased in the initial rounds and stabilized afterward without significantly decreasing over the interaction for lexical and gestural alignment. The results are consistent with the prediction that people maintain alignment in words and iconic gestures after the emergence of lexical and gestural alignment, although the interlocutors could have dropped gestural alignment after the establishment of shared lexical expressions. This suggests that when people communicate about novel referents for which they do not share conventionalized labels, they effectively use speech and gestures to jointly create shared multimodal labels and keep aligning in words and gestures whenever they communicate about the referents.

Although our study suggests that the process of establishing and maintaining shared labels is multimodal in nature, more fine-grained work is necessary to confirm the hypothesis that people establish and maintain the shared expressions multimodally in joint meaning-making processes. In particular, to study *how* shared expressions are calibrated and maintained over time, we must track the contents of the lexical and gestural expressions for each instance of alignment in future work. This might reveal interactants' use of "multimodal conceptual pacts".

RQ2: How are alignment in speech and gestures related over time?

As for the second research question, we found a positive association between lexical alignment rate and gestural alignment rate, supporting the prediction that alignment in speech and co-speech gestures coincide. The result can potentially be explained by the interactive alignment model (Pickering & Garrod, 2004), which proposes that alignment at one linguistic representation level (e.g., the lexical level) leads to alignment at other linguistic representation levels (e.g., gesture).

However, it is important to acknowledge that our study was not designed to test the interactive alignment model and its extension to co-speech gestures. Also, as illustrated in Figure 4, lexical and gestural alignment often coincide in the same interactive sequence, rather than alignment in words leading to alignment in gestures or vice versa. Thus, further research is needed to deepen our understanding of the mechanisms underpinning the relationships between alignment in speech and gestures.

Conclusion

The present study examined people's use of behavioral alignment in establishing shared labels for novel referents and the association between lexical and gestural alignment. Our findings corroborate earlier work showing that people deploy speech and gesture to establish multimodal shared labels (Holler & Wilkin, 2011; Macuch Silva et al., 2020; Rasenberg et al., 2022). Furthermore, our results show that lexical and gestural alignment go hand-in-hand and that alignment is not only used in the initial phase, but throughout the interaction. The present study also showed that the amount of lexical and gestural alignment is positively associated, suggesting a close link between alignment in speech and gesture.

[3]The linear pattern in the scatterplot is plausibly a byproduct of the calculation of the alignment rate, which was done by dividing the amount of alignment by the trial duration. For example, suppose that the amount of alignment is 10 for lexical and gestural alignment for three different trials whose duration was 20, 40, and 60 seconds. In that case, the alignment rate will be 3.3, 6.6, and 10, which creates a linear pattern.

Acknowledgments

We thank Dr. Mark Dingemanse for providing feedback on the draft. This research was funded by the Max Planck Society for the Advancement of Science (www.mpg.de/en).

References

Barry, T., Griffith, J., De Rossi, S., & Hermans, D. (2014). Meet the Fribbles: Novel stimuli for use within behavioural research. *Frontiers in Psychology*, *5*.

Bartolozzi, F., Jongman, S. R., & Meyer, A. S. (2021). Concurrent speech planning does not eliminate repetition priming from spoken words: Evidence from linguistic dual-tasking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(3), 466–480. https://doi.org/10.1037/xlm0000944

Bergmann, K., & Kopp, S. (2012). Gestural alignment in natural dialogue. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34).

Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, *42*(9), 2355–2368. https://doi.org/10.1016/j.pragma.2009.12.012

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

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400.

Chen, S.-Y., Feng, Z., & Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *Journal of Thoracic Disease*, *9*(6), 1725–1729. https://doi.org/10.21037/jtd.2017.05.34

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association. https://doi.org/10.1037/10096-006

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

Couper-Kuhlen, E., & Selting, M. (2018). *Interactional linguistics: Studying language in social interaction*. Cambridge University Press. https://doi.org/10.1007/9781139507318

Crible, L., Gandolfi, G., & Pickering, M. J. (2024). Feedback quality and divided attention: Exploring commentaries on alignment in task-oriented dialogue. *Language and Cognition*, 1–29. https://doi.org/10.1017/langcog.2023.65

Dideriksen, C., Christiansen, M. H., Tylén, K., Dingemanse, M., & Fusaroli, R. (2022). Quantifying the interplay of conversational devices in building mutual understanding. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001301

Eijk, L., Rasenberg, M., Arnese, F., Blokpoel, M., Dingemanse, M., Doeller, C. F., Ernestus, M., Holler, J., Milivojevic, B., Özyürek, A., Pouw, W., van Rooij, I., Schriefers, H., Toni, I., Trujillo, J., & Bögels, S. (2022). The CABB dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage*, *264*, 119734. https://doi.org/10.1016/j.neuroimage.2022.119734

Ghaleb, E., Rasenberg, M., Pouw, W., Toni, I., Holler, J., Özyürek, A., & Fernández, R. (2024). Analysing cross-speaker convergence through the lens of automatically detected shared linguistic constructions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(46).

Green, C., & Sun, H. (2021). Global estimates of syntactic alignment in adult and child utterances during interaction: NLP estimates based on multiple corpora. *Language Sciences*, *85*, 101353. https://doi.org/10.1016/j.langsci.2020.101353

Hartsuiker, R. J., Bernolet, S., Schoonbaert, S., Speybroeck, S., & Vanderelst, D. (2008). Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, *58*(2), 214–238. https://doi.org/10.1016/j.jml.2007.07.003

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, *44*(6), e12845. https://doi.org/10.1111/cogs.12845

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

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python.

Ivanova, I., Horton, W. S., Swets, B., Kleinman, D., & Ferreira, V. S. (2020). Structural alignment in dialogue and monologue (and what attention may have to do with it). *Journal of Memory and Language*, *110*, 104052. https://doi.org/10.1016/j.jml.2019.104052

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

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

Kimbara, I. (2008). Gesture Form Convergence in Joint Description. *Journal of Nonverbal Behavior*, *32*(2), 123–131. https://doi.org/10.1007/s10919-007-0044-4

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

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology*, *14*(1), 78–106. https://doi.org/10.1016/0010-0285(82)90005-6

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. (2005). *Gesture and thought*. University of Chicago Press.

Norrick, N. R. (1987). Functions of repetition in conversation. *Text - Interdisciplinary Journal for the Study of Discourse*, *7*(3), 245–264. https://doi.org/10.1515/text.1.1987.7.3.245

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

Özyürek, A. (2018). Role of Gesture in Language Processing: Toward a unified account for production and comprehension. In *The Oxford Handbook of Psycholinguistics* (pp. 591–607). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198786825.013.25

Özyürek, A., & Woll, B. (2019). Language in the visual modality: Co-speech gesture and sign language. In *Human language: From genes and brain to behavior* (pp. 67–83). MIT Press.

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

Pickering, M. J., & Garrod, S. (2021). *Understanding dialogue: Language use and social interaction* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108610728

Põldvere, N., Johansson, V., & Paradis, C. (2021). Resonance in dialogue: The interplay between intersubjective motivations and cognitive facilitation. *Language and Cognition*, *13*(4), 643–669. https://doi.org/10.1017/langcog.2021.16

R Core Team. (2020). R: A language and environment for statistical computing.

Rasenberg, M. (2023). *Mutual understanding from a multimodal and interactional perspective* (Doctoral dissertation). Radboud University. Nijmegen, Netherlands.

Rasenberg, M., Özyürek, A., Bögels, S., & Dingemanse, M. (2022). The primacy of multimodal alignment in converging on shared symbols for novel referents. *Discourse Processes*, *59*(3), 209–236. https://doi.org/10.1080/0163853X.2021.1992235

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

Rubin, M. (2021). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. *Synthese*, *199*(3), 10969–11000. https://doi.org/10.1007/s11229-021-03276-4

Sinclair, A., & Fernández, R. (2021). Construction coordination in first and second language acquisition. *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*.

Winter, B., & Bürkner, P.-C. (2021). Poisson regression for linguists: A tutorial introduction to modelling count data with brms. *Language and Linguistics Compass*, *15*(11), e12439. https://doi.org/10.1111/lnc3.12439

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.