Conceptual alignment in a joint picture-naming task performed with a social robot

Giusy Cirillo, Elin Runnqvist, Kristof Strijkers, Noël Nguyen, Cristina Baus

Conceptual alignment in a joint picture-naming task performed with a social robot

Giusy Cirillo[1,2], Elin Runnqvist[1,2], Kristof Strijkers[1,2], Noël Nguyen[1,2], Cristina Baus[3]

1 Laboratoire Parole et Langage (LPL), CNRS & Aix-Marseille University, Aix-en-Provence, France
2 Institute for Language, Communication and the Brain (ILCB), Marseille, France
3 Department of Cognition, Development and Educational Psychology, University of Barcelona, Spain

Abstract

Conversation entails a tight coordination between the interlocutors in terms of co-representation and linguistic alignment (e.g., word choices). In this study we investigated whether people conceptually align in a language task with a robot. 24 French native speakers alternated with an artificial partner in naming images of objects belonging to different semantic categories (e.g., mammals, clothes…). For five out of fifteen categories the robot produced the item's category instead of the preferred basic-level name. Logistic regression models on participants' errors revealed that they adapted to the robot's conceptual choices, and produced more category names over the course of the experiment. This pattern was most prominent for the semantic categories for which the robot had used a category name, and importantly, it applied to novel items. These results provide strong evidence for conceptual alignment affecting word choices, indicating that prototypical concepts can be overwritten in conversation to adapt to the interlocutor.

Keywords: joint action; spoken word production; conceptual alignment; lexical alignment; co-representation; artificial partner; picture naming

Introduction

People engage in interactions daily. When playing or dancing together and when talking to each other, individuals transmit and react to relevant information from their partner with the aim of making their performance smoother and faster. These 'joint actions' are a fundamental part of social cognition, as they explain not only how humans' social bonds are established, but also how they mutate depending on the situation and the partner. An intrinsic characteristic of any joint action is alignment (Pickering & Garrod, 2007). Activity-partners align their action representations at different levels (motor, cognitive) and this is accomplished through prediction mechanisms (Sebanz et al., 2006; Sebanz & Knoblich, 2009). That is, prediction of others' actions involves processes engaged in the planning and performance of one's own actions and this is thought to be accomplished by the activation of forward models (imitative plans) in the motor system (Pesquita et al., 2018).

In the context of language, alignment has been described as crucial for successful communication (Pickering & Garrod, 2013; Pickering & Garrod, 2006). Accumulated evidence has revealed that alignment occurs because speakers prime each other at different levels of representation (e.g., lexical, syntactic, etc.). Numerous studies show that speakers mimic a number of non-verbal behaviors, including facial expressions (Dimberg et al., 2000), limb movements (Kilner et al., 2003), gestures (Bergmann & Kopp, 2012; Louwerse et al., 2012), and posture (Shockley et al., 2007), as well as verbal behaviors. For example, people align verbally with their interlocutors in terms of accent and speech rate and other phonetic dimensions (Giles et al., 1991; Pardo, 2006), sentence structure (Branigan et al., 2000) and word choices (Brennan & Clark, 1996; Garrod & Pickering, 2004). Particularly convincing for the notion of alignment, is that speakers even copy atypical lexical responses such as rarely used synonyms (e.g., Brennan & Clark, 1996) and infrequent syntactic structures such as passives (e.g., Bock, 1986). Findings like these clearly show that linguistic alignment is a

powerful communicative mechanism, capable even of overriding more frequent verbal behaviors.

The above work shows that speakers align to the utterances of their partners, even to atypical ones. In the present study we wanted to explore linguistic alignment beyond the copying of verbal utterances per se and asked whether a speaker's lexical choices would be affected by the alignment with their partner's conceptual knowledge. That is, whether speakers adopt their partner's conceptual patterns leading to the production of infrequent, yet meaningful lexico-semantic choices, and whether it generalizes to all lexical items belonging to that conceptual category (e.g., saying a category name instead of the object name for all objects of a given category). To do so, we adopted a joint-action design.

Joint action settings in which two people share a linguistic task have been used to study parity of lexical representations between speakers and listeners and predictive mechanisms. In particular, joint picture-naming tasks have been employed to explore whether lexical representations are shared across speakers and used by each speaker to predict the interlocutor's upcoming utterance (e.g., Baus et al., 2014; Gambi et al., 2015). These designs partly reproduce the dynamics of turn-taking in conversation by having two participants naming pictures in an alternate manner. Most of those studies have focused on the interference and facilitation effects observed in participants as a result of naming objects with a partner, compared with performing the task individually. For instance, subjects' naming latencies have been pointed out to be slower when participants are made to believe that their partner is concurrently naming the item (Gambi et al., 2015). In addition, the cumulative semantic effect (i.e., a slowdown in naming latencies after a sequence of semantically-related pictures) previously shown in individual naming tasks by Howard et al. (2006), has also been found when dyads of participants performed the task in an alternate way, which suggests that speakers use shared lexical representations (Kuhlen & Abdel Rahman, 2017).

In another picture-naming study by Baus et al. (2014), representation was investigated in an ERP experiment, with a set up that comprised go (i.e., subject naming the picture), other-go (i.e., partner naming the picture) and no-go(nobody naming the picture) trials. In all experimental naming conditions, EEG waveforms for high frequency words were compared with those for low frequency words (i.e., the frequency effect, Almeida et al., 2007; Navarrete et al., 2006; Oldfield & Wingfield, 1965; Strijkers et al., 2010). Authors found comparable ERP modulations as a function of the lexical frequency effect independently of who was speaking and who was listening, demonstrating that participants predicted the lexical representations of their interlocutors. Overall, all those studies are relevant in revealing the appropriateness of joint action settings to explore the dynamics of lexical co-representation and predictive mechanisms.

In the current study joint action constituted the *method* of investigation, while linguistic (lexical and conceptual) alignment was the *object* of investigation. We asked participants to perform a picture-naming task together with a social robot — Furhat Robot. We made this methodological choice in order to accurately control the dynamics of the human participant and robot's joint performance, and to easily manipulate the robot's lexical choices. Indeed, we manipulated the response of the robot for a regular subset of trials, in which the robot did not give the basic level name of the item but its semantic category name instead (e.g., *fruit* = fruit instead of *pêche* = peach*)*. As people naturally tend to use the basic level name, a phenomenon known as basic-level advantage (Rogers & Patterson, 2007), we expected that if co-representation was to take place it was to be found in a progressive adaptation to the behavioral patterns of the robot, evident by the use of the robot's conceptual choices.

We aimed to show a naming pattern going beyond repetition of the same word in lexical alignment and address the capacity of speakers to adapt to the conceptual language space of the robot. Indeed, in our experiment robot and participants did not share the same items but the semantic category only, so that the use of a misreferred name is to be interpreted as adaptation at the conceptual level, especially in the case when a category name is used to items belonging to the categories not previously named with the new label by the robot. In short, this would mean that interlocutors align conceptually rather than to simple lexical choices. We believe that our data can contribute to understanding how conceptual adaptation affects lexical choices, and the time dynamics of such conceptually driven alignment.

Methods

Participants

Twenty-four participants (5 men; age: *M* = 22.25 years, *SD* = 2.9, range = 18–30 years) participated in the study. All participants were right-handed, native speakers of French with normal or corrected-to-normal vision. None of the participants reported any neurological disorders, psychiatric disorders, or speech/language impairments. They received 10 euros for their participation. Given the novelty of our hypothesis, the choice of the sample size was considered reasonable on the basis of previous literature on lexical alignment over atypical names. In particular, in a series of experiments, Branigan et al. (2011) found that subjects started to adopt a misreferred name for pictures previously named with the same atypical name by their artificial partner. While their specific research question was centered on belief about their partner (whether a human, a computer, or a very sophisticated or unsophisticated machine), they were able to find a significant effect of disfavored name prime on the adoption of misreferred names in speech using 9 x 2 (name prime) experimental items for a pool of 32 participants. The proportion of disfavored target responses participants gave when interacting with a computer was .06 after a favored name prime and .77 after a disfavored name prime.

In our experiment, participants were asked to name 225 pictures in total, out of which 75 belonged to the categories included in the category naming level condition of the robot, and 150 in the basic naming level condition. The number of data-points we collected for 24 participants was sufficient to have a considerable effect for our atypical response pattern, especially considering that participants and robot never shared the same pictures, but only the superordinate semantic category

Furhat Robot

In our experiment we used Furhat (https://www.furhatrobotics.com/; Al Moubayed et al., 2012), a humanoid robot equipped with a sophisticated back-projection system with a 3D-printed mask which can resemble anyone's face. In particular, its set up was built to control its facial expressions (including eyebrows, mouth, lips) to make its movements smooth and natural. The robot can perform two of the main facial gestures judged to be fundamental in interactional conversation, namely lip synchronization and gaze direction (Moubayed et al., 2013).

The felicitous use of Furhat for a natural joint-activity setting has been demonstrated in a number of experiments, where the robot has been employed as partner in both perception and production tasks (Moubayed et al., 2013; Rauchbauer Birgit et al., 2019; Skantze, 2016).

Materials

The set of visual stimuli consisted of 450 pictures. 377 of them were taken from the dataset MultiPic (Duñabeitia et al., 2018). The remaining 73 pictures were drawn by a professional designer who used the MultiPic picture format as a model. Pictures were chosen on the basis of the semantic category and word frequency of their corresponding lexical label.

The set of pictures included 15 semantic categories (e.g., fruits, mammals, clothes…), with 30 items in each (see Appendix A). Concept typicality (Morrow & Duffy, 2005; Woollams, 2012) for each name was assessed via a questionnaire, in which 30 students of Aix-Marseille University rated how much each word was representative for the semantic category indicated in brackets in a 5-point Likert scale, where 1 corresponded to 'not at all (representative)' and 5 to 'very much (representative)'. Only items that had a mean of at least 3.0 were selected for the main experiment. The same pool of participants carried out another questionnaire on subjective frequency, in which they had to evaluate how frequently they encountered a given name in a 5-point Likert scale, where 1 corresponded to 'very rarely' and 5 to 'very often'.

Afterwards, the frequency value for each item was extracted from a French word database (Lexique, http://www.lexique.org/, New et al., 2001). We averaged the means of the frequency values for words contained in books and words contained in film subtitles, to account for both spoken and written usage. A median split of the stimuli resulted in a mean frequency value for the high frequency items of 72.5 occurrences per million (sd = 108.2, subjective frequency: mean = 3.3, sd= 0.8) and a mean frequency value for low frequency items of 5.7 occurrences per million (sd = 3.9; subjective frequency: mean 2.3, sd= 0.8). The lexical frequency was taken as a measure of the ease in lexical access, in line with previous work in production using picture-naming (e.g., Baus et al., 2014).

Half of the items of each semantic category (15 items x 15 semantic categories = 225 items) were assigned to the subject's trials (go trials) and the other half to the robot's trials (other-go trials). In this way, the participant and the robot shared the same semantic categories, while they had to name different pictures. In addition, for both go and other-go trials, half of the trials were high-frequency items and the other half low-frequency items. Pictures were presented within a square, of which the color indicated whether it was the human participant's or the robot's turn to speak. The assignment of colors (blue and green) to naming condition was counterbalanced across participants.

The robot's responses were pre-recorded using the synthesized voice of Furhat Robotics and time locked to the picture presentation in the experiment. They consisted of 225 productions of unique monosyllabic, disyllabic and trisyllabic words of between 200 ms and 1000 ms of duration each and were played via the loudspeakers of the robot. We created six sets of randomized naming latencies (M= 930 ms, SD=250 ms, range= 500-1600 ms) to mimic the intra-individual variability in naming latency observed in human speakers. The numbers were calculated after asking 20 students from Aix-Marseille University to name the pictures and recording their naming latencies. The robot was programmed to produce the basic-level name (e.g., *main* = hand, *castagne* = chestnut) for items belonging to 10 semantic categories, and the semantic category name (e.g., *mammifère* = mammal, *outil* = tool) for items corresponding to the remaining five semantic categories.

Categories and items as well as the robot's reaction time sets were counterbalanced across six lists following a Latin square design in order to avoid item, category, or robot's naming

latencies biases. Each list comprised a basic naming condition and a category naming condition, as related to the response pattern of the robot. Each participant was exposed to one list.

Procedure

At the beginning of the study, participants were introduced to Furhat, and given 5 minutes to freely conversate with the robot. After the introduction, participants were told that they would complete a joint picture-naming task with the robot, in which they had to name as quickly as possible the pictures presented within the square of the color to which they were assigned and remain silent for the rest of the trials to let Furhat speak.

Participant and robot were positioned alongside facing the same computer screen, on which the visual stimuli were displayed (see Fig. 1). Each trial started with the presentation of a fixation cross (for 500 ms), followed by the picture with the color cue corresponding to go or other-go trials (for 3000 ms) and by a white screen (for 500 ms) to separate the trials (see Fig. 2). Every 90 trials, participants were asked to take a 2-minute pause.

Verbal responses and naming latencies were recorded for each participant. Response times of the robot were pre-set and responses were played during Furhat's turn via the loudspeakers of the robot, and included basic-level names (for the basic naming level condition) and semantic category names (for the category naming level condition).

At the end of the experiment, participants filled out a questionnaire where they rated on 5-point Likert scales a) their subjective view on artificial agents as well as their familiarity with humanoid robots, b) the quality of the experiment, Furhat's performance, and their ability to coordinate with the robot, c) the physical and social characteristics of Furhat (appearance, facial expression, voice, behavior, and sociability).

Fig. 1. Experimental setting.

Participant and robot are positioned alongside and face the same computer screen, where the pictures are shown.

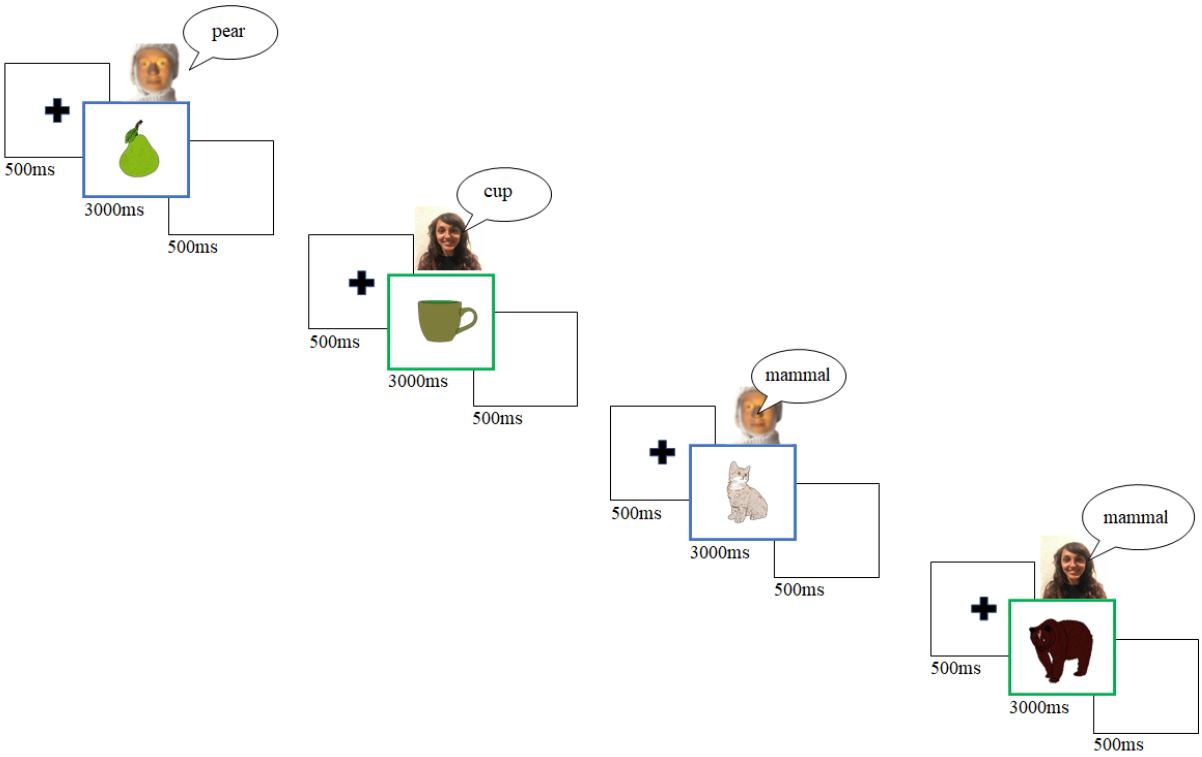

Fig. 2. Example of the timeline of the experiment.

In the first trial (on the top left), the robot produces the basic-level name of the object (basic condition), and the participant in the second trial does the same. In the third trial the robot produces the category name (category condition). In the fourth trial the participant uses the category name as well, as the item belongs to the same category. This is an example of response suggesting conceptual alignment within category.

Data analysis

We performed three types of analysis on the participants' behavioral responses (go trials): a naming latency analysis, an analysis of error rates and an analysis aimed to compare the effect of block/order of trials and naming level condition (basic vs category) over two types of alternative responses participants used to name the pictures, namely semantic category names versus semantically related names.

Naming latencies and error rates analyses were carried out by fitting Generalized Linear Mixed Effects models using the lmr4 package in R (Bates et al., 2015), where frequency (low vs high), naming level condition (basic vs category) and block (1 to 4, corresponding to the items named in between the two minute pauses) were taken as fixed factors while item and participant were taken as random factors. Moreover, we used the quantitative variable 'order of trials' within an additional analysis of the alternative responses in order to have a more continuous index of time. All variables were contrast-coded using the Helmert contrast method, in which each level of a factor is contrasted to the mean of the previous ones. This method was implemented for the variable 'block', as we were able to compare each time point to the average of the previous ones.

Missing responses, hesitations and verbal responses which differed from the target name were excluded from the naming latencies analysis (N = 1000). In addition, values that were 2.5 standard deviations above or below the mean response time for high and low frequency words respectively were excluded (N= 117). Naming latencies were log-transformed for a better fit of the model.

Errors (N = 1000), and alternative responses were processed by fitting logistic regression models, more suitable to binary data. From the different types of alternative responses - excluded from the naming latencies analysis, included in the error rates analysis- we selected the two largest groups, namely the semantic category names (e.g., *tool* for *hammer*; N = 176) and the names belonging to the same semantic category of the target name (e.g., *bracelet* for *necklace*; N = 218). These response variables were subject to separate analysis. In particular, we fitted two models per type of response, as including word frequency, naming level and block together would not allow the model to converge. We therefore had a first model including the type of alternative response as dependent variable (i.e., category name or semantic-related name) with frequency and block as fixed factors, and a second model including naming level and block as fixed factors. Similarly, we followed the same procedure using the continuous variable 'order of trials'. Via this procedure we aimed to test whether co-representation was mirrored in the response of the participant, in particular by looking at the distribution of the semantic category names across time and naming level condition. The semantically related names were chosen as a control group since we did not expect an increase in the use of these alternative names over the course of the experiment.

Finally, we performed a series of Spearman's rank correlations between groups of Likert-scale questions and number of category responses. We then took the most representative questions per group to see whether lexical entrainment could be explained by certain beliefs about robots or whether category responses would be corroborated by stated awareness of one's adaptive/predictive behavior.

Results

The naming latency results revealed that low frequency words were produced significantly later than high frequency words (Mean HF = 1148 ms and SD = 299 ms; Mean LF = 1241 ms and SD = 328 ms; $b = 0.05144$, s.e. = 0.006935, $t = 7.418$, $p < 0.001$). Moreover, participants responded faster towards the end of the experiment, and in particular in the fourth block as compared to the previous blocks ($b = -0.03287$, s.e. = 0.002135, $t = -15.398$ $p < 0.001$). However, there was no effect of naming level condition ($p = 0.361$) nor of any interaction ($p > 0.05$), showing that the frequency effect was overall distributed in the same way across both conditions.

The error rates showed a similar frequency effect, with low frequency words eliciting errors more often as compared to high frequency words (Prop. HF = 0.1 and SD = 0.3; Prop LF = 0.26 and SD = 0.4; $b = 0.71$, s.e. = 0.09, $z = 7.36$, $p < 0.001$). However, the model did not reveal any effect of the robot's response condition ($p = 0.715$), nor block ($p > 0.05$) nor an interaction between these ($p > 0.05$).

The model exploring category responses across time through the continuous variable 'order of trials' revealed a significant effect of this variable ($b = 0.005634$, s.e. = 0.001832, $z = 3.076$, $p = 0.002$), indicating that participants were more likely to produce category responses towards the end of the experiment. A similar finding was reproduced using the variable block (1:4) as an indicator of time, where a significant effect of block 4 contrasted to the average of the previous ones resulted from the analysis ($b = 0.17298$, s.e. = 0.06225, $z = 2.779$, $p = 0.005$).

The model revealed that category names were produced more often in the category level condition ($b = 0.42575$, s.e. = 0.10594, $z = 4.019$, $p < 0.001$; see Fig. 3 for the effect of naming level and block on category responses), and for low frequency items ($b = 0.8277$, s.e. = 0.1661, $z = 4.982$, $p < 0.001$). However, we did not find any interaction among the variables ($p > 0.05$).

We followed the same procedure to analyze semantically related names and found an inverse effect of the continuous variable order (b = -0.0043196, s.e. = 0.0014741, z = -2.930, p = 0.003) and of the variable block (b = -0.12802, s.e. = 0.05838, z = -2.193, p = 0.023), indicating that names belonging to the same semantic category of the target were named less often towards the end of the experiment (see Fig. 4). In addition, the model revealed the same frequency effect (b = 0.61380, s.e. = 0.15304, z = 4.011, p < 0.001), but no effect for naming level condition (p = 0.212), nor any interaction (p > 0.05).

From the analysis on the responses of the questionnaire, we found that the number of category productions was negatively correlated with familiarity of virtual and physical robots (representative question: *I am familiar with humanoid robots (virtual or physical);* r = -0.445, p = 0.029) and imageability (representative question: *I can easily imagine to interact with robots*) = -0.5, p < 0.012), and positively correlated with statement of adaptive prediction (representative question: *My way of predicting changed over the course of the experiment*; r = 0.57, p = 0.003). In contrast, we did not find any effect related to a stated difficulty of the task/experiment (representative question: *I found the experiment difficult*; r = -0.15, p = 0.341) nor to the human-like appearance of Furhat (e.g., *Rate how human-like the facial expressions are*, r = 0.147, p > 0.492). Finally, subjects rated Furhat as very human-like in terms of his appearance (79%), his sociability (79%) and his facial expression (58%) but judged as less human his voice (42%) and his behavior (33%).

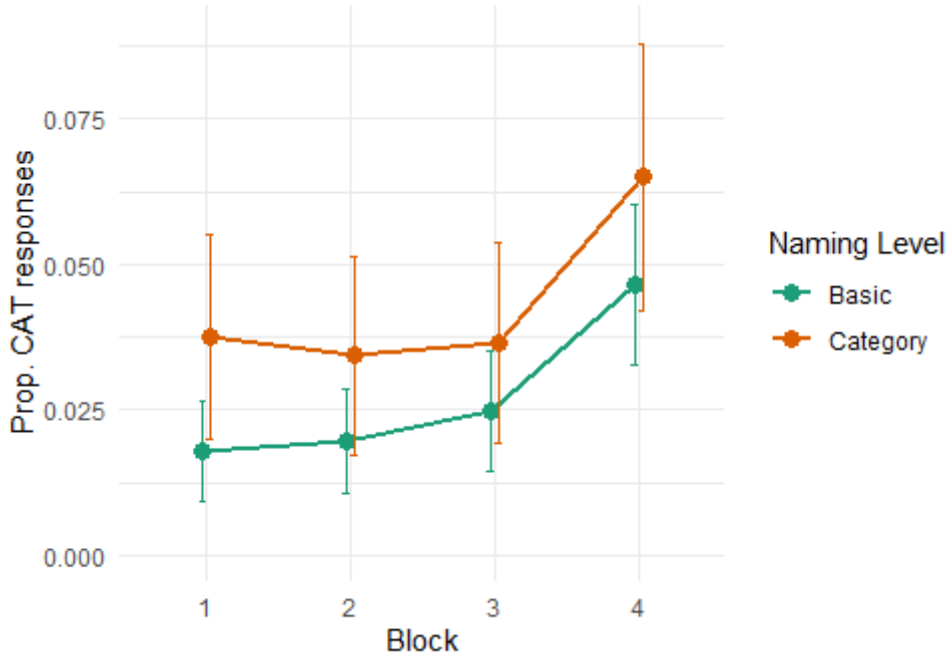

Fig. 3: **Category responses by block and naming level**.

Mean and confidence intervals (95%) for category responses by block (1-4) and naming level (basic vs category).

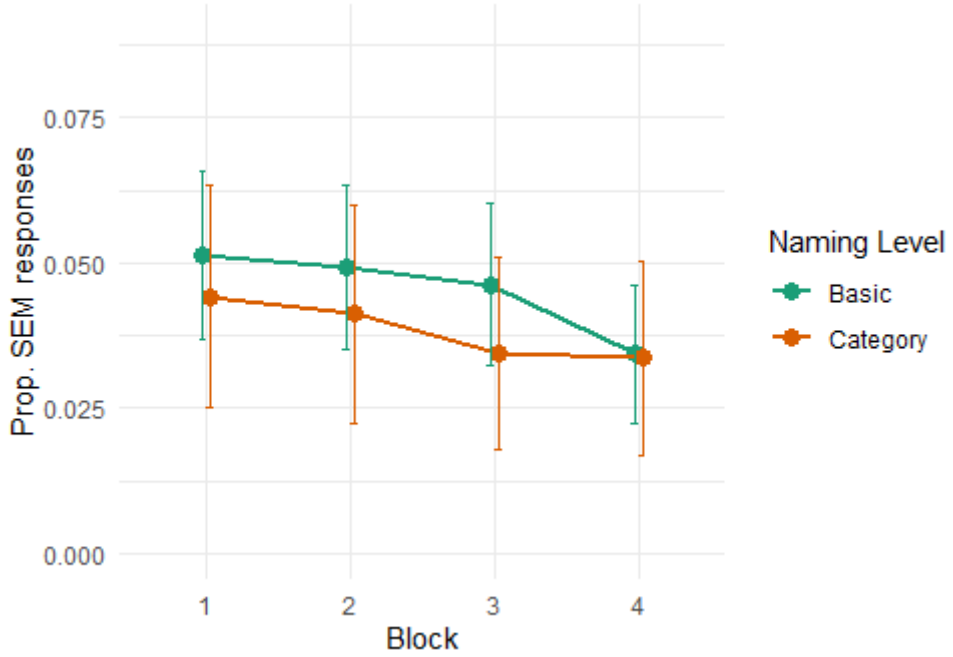

Fig. 4: **Semantic-related responses by block and naming level**.

Mean and confidence intervals for semantic-related responses by block (1-4) and naming level (basic vs category)

Discussion

In this study, we explored whether people align to the conceptual choice patterns of an artificial partner in a joint naming task. The robot was programmed to produce the semantic category name of objects belonging to 5 semantic categories (category condition), while for the rest of the trials it produced the basic level name (basic-level condition). By doing this we were able to create an atypical, yet meaningful lexico-semantic choice pattern.

Our results showed that participants produced more frequently these atypical category names instead of the preferred basic-level names when naming pictures belonging to the same semantic category as those pictures that were given a category name by the robot. This data pattern fits previous demonstrations where through alignment speakers repeat infrequent lexical names or structures uttered by a human or computer partner (Bock, 1986; Brennan & Clark, 1996; Ivanova et al., 2012). Importantly, our results go beyond the direct emulation of an atypical lexical label, and show a conceptual alignment modifying an interlocutor's verbal behavior for previously unseen pictures. That is, we observed that the robot's category response (e.g., naming the picture of a hammer as 'tool') caused the human interlocutor to utter a category name for new images (e.g., naming the previously not seen picture of a 'nail' as 'tool') . By performing fine-grained analyses over the course of the experiment for these category responses of the participant, we furthermore observed that this form of conceptual alignment is subject to two temporal dynamics: (1) It emerged from the first block, indicating very rapid, possibly automatic, adaptation to the robot's concepts; (2) There was a significant increase in the participants' alignment in the final block, indicating an additive effect of gradual, possibly more conscious alignment with the robot's concepts. The latter is also supported by the ratings participants gave at the end of the experiment, where they reported that they were trying to predict the robot's behavior, suggesting that participants were to some extent aware of the robot's lexico-semantic response pattern.

Importantly, the current observations of conceptual alignment cannot be accounted for by assuming that participants simply strategically updated their type of response nor that they were perhaps confused by the robot's replies of what the actual task was. Indeed, a simple explanation to this behavior would be that the category responses given by the participants in this experiment are not an index of proper linguistic alignment, but rather that category responses were equally valid, and participants gave those responses when they did not know the basic level name of a presented object. However, several observations make this alternative account improbable: First, as mentioned, the speed with which participants engaged in conceptual alignment suggests it also contains an automatic component. Second, the fact that the conceptual alignment was most pronounced for those categories where the robot gave a category response, suggests that conceptual alignment is contextual. And third, when we compare with other alternative responses given by the participants, namely members belonging to the same semantic category (semantic errors), we find the reverse pattern. That is, there were more semantic errors in the beginning of the experiment compared to the final two blocks, and these types of errors were similar for the basic-level and category-level condition. Clearly, if our results were merely due to participants' strategy of giving a category name when uncertainty or confusion about the basic level name was high, the prediction would be to see more category responses in the first two blocks (where semantic confusion and errors were higher), rather than in the last part of the experiment.

In addition, another intriguing observation was that participants also started giving category names which were different from those produced by the robot. Though it was significantly less compared to those items which did belong to the same semantic category where the robot gave a category label, this generalization behavior of the participants most likely reflects spreading activation of conceptual choices. That is, due to the participants adaption to category-level lexical concepts, on some occasions this adaptation may spill over to other items (and the more conceptual adaption, the higher probability for these spill-over effects, as evidenced in the final block). More importantly for our purposes, this generalization behavior is interesting because it additionally confirms the interpretation that the current data originate at the conceptual level of processing, rather than any other form of alignment or task-dependent strategies.

In sum, the current results are best understood in the light of the theory of linguistic alignment and the link between language production and perception, in terms of how the perception of a linguistic behavior modulates cognitive states and affects production (Marsh et al., 2009; Pickering & Garrod, 2013). In this manner, our results support the notion that listeners update their speech production in line with their partner's response patterns, and that they do that not simply by choosing to repeat the same lexical names, but by adopting the same conceptual word knowledge of the interlocutor. The behavioral shift of our participants who started imitating the response pattern of their artificial interlocutor can be understood as a way to connect to the robot, to bring about the same 'change in the environment' as their partner (Sebanz et al., 2006). Our results contribute and extend the theory of alignment in that this behavioral shift was characterized by a conceptual adaptation of the lexico-semantic patterns of the interlocutor.

Author Contributions

All authors developed the study concept and contributed to the study design. Testing, data collection and data analysis were performed by G. C. under the supervision of all authors. G.C. drafted the manuscript, and E.R., K.S., N.N and C.B. provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgement

This study has received funding from the "Investissements d'Avenir" French Government program managed by the French National Research Agency (reference : ANR-16-CONV-000X / ANR -17-EURE-00XX) and from Excellence Initiative of Aix-Marseille University - A*MIDEX" through the Institute of Language, Communication and the Brain. Giusy Cirillo was supported by the Ecole Doctorale Cognition, Langage, Education / ED 356 of Aix-Marseille University. Cristina Baus was supported by the Ramon y Cajal research program (RYC2018-026174-I). E.R. has benefited from support from the French government, managed by the French National Agency for Research (ANR) through a research grant (ANR-18- CE28-0013). K.S. was supported by a research grant of the ANR (ANR-16-CE28-0007-01). We are grateful to Laurent Prévot for useful discussions and the technical help with the robot Furhat.

Bibliography

Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (2012). Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction. In A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, & V. C. Müller (Eds.), *Cognitive Behavioural Systems* (pp. 114–130). Springer. https://doi.org/10.1007/978-3-642-34584-5_9

Almeida, J., Knobel, M., Finkbeiner, M., & Caramazza, A. (2007). The locus of the frequency effect in picture naming: When recognizing is not enough. *Psychonomic Bulletin & Review*, *14*(6), 1177–1182. https://doi.org/10.3758/BF03193109

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baus, C., Sebanz, N., Fuente, V. de la, Branzi, F. M., Martin, C., & Costa, A. (2014). On predicting others' words: Electrophysiological evidence of prediction in speech production. *Cognition*, *133*(2), 395–407. https://doi.org/10.1016/j.cognition.2014.07.006

Bergmann, K., & Kopp, S. (2012). Gestural Alignment in Natural Dialogue. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34). https://escholarship.org/uc/item/73z0q063

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387. https://doi.org/10.1016/0010-0285(86)90004-6

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

Dimberg, U., Thunberg, M., & Elmehed, K. (2000). Unconscious Facial Reactions to Emotional Facial Expressions. *Psychological Science*, *11*(1), 86–89. https://doi.org/10.1111/1467-9280.00221

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808–816. https://doi.org/10.1080/17470218.2017.1310261

Gambi, C., Van de Cavey, J., & Pickering, M. J. (2015). Interference in joint picture naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 1–21. https://doi.org/10.1037/a0037438

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11. https://doi.org/10.1016/j.tics.2003.10.016

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: Developments in applied sociolinguistics* (pp. 1–68). Editions de la Maison des Sciences de l'Homme. https://doi.org/10.1017/CBO9780511663673.001

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. *Cognition*, *100*(3), 464–482. https://doi.org/10.1016/j.cognition.2005.02.006

Ivanova, I., Pickering, M. J., McLean, J. F., Costa, A., & Branigan, H. P. (2012). How do people produce ungrammatical utterances? *Journal of Memory and Language*, *67*(3), 355–370. https://doi.org/10.1016/j.jml.2012.06.003

Kilner, J. M., Paulignan, Y., & Blakemore, S. J. (2003). An Interference Effect of Observed Biological Movement on Action. *Current Biology*, *13*(6), 522–525. https://doi.org/10.1016/S0960-9822(03)00165-9

Kuhlen, A. K., & Abdel Rahman, R. (2017). Having a task partner affects lexical retrieval: Spoken word production in shared task settings. *Cognition*, *166*, 94–106. https://doi.org/10.1016/j.cognition.2017.05.024

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

Marsh, K. L., Richardson, M. J., & Schmidt, R. C. (2009). Social Connection Through Joint Action and Interpersonal Coordination. *Topics in Cognitive Science*, *1*(2), 320–339. https://doi.org/10.1111/j.1756-8765.2009.01022.x

Morrow, L. I., & Duffy, M. F. (2005). The representation of ontological category concepts as affected by healthy aging: Normative data and theoretical implications. *Behavior Research Methods*, *37*(4), 608–625. https://doi.org/10.3758/BF03192731

Moubayed, S. A., Skantze, G., & Beskow, J. (2013). The furhat back-projected humanoid head–lip reading, gaze and multi-party interaction. *International Journal of Humanoid Robotics*, *10*(01), 1350005. https://doi.org/10.1142/S0219843613500059

Navarrete, E., Basagni, B., Alario, F.-X., & Costa, A. (2006). Does word frequency affect lexical selection in speech production? *Quarterly Journal of Experimental Psychology*, *59*(10), 1681–1690. https://doi.org/10.1080/17470210600750558

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE^{TM}//A lexical database for contemporary french : LEXIQUE^{TM}. *Annee Psychologique - ANNEE PSYCHOL*, *101*, 447–462. https://doi.org/10.3406/psy.2001.1341

Oldfield, R. C., & Wingfield, A. (1965). Response Latencies in Naming Objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281. https://doi.org/10.1080/17470216508416445

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393. https://doi.org/10.1121/1.2178720

Pesquita, A., Whitwell, R. L., & Enns, J. T. (2018). Predictive joint-action model: A hierarchical predictive approach to human cooperation. *Psychonomic Bulletin & Review*, *25*(5), 1751–1769. https://doi.org/10.3758/s13423-017-1393-6

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(04), 329–347. https://doi.org/10.1017/S0140525X12001495

Pickering, Martin J., & Garrod, S. (2006). Alignment as the Basis for Successful Communication. *Research on Language and Computation*, *4*(2), 203–228. https://doi.org/10.1007/s11168-006-9004-0

Pickering, Martin J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110. https://doi.org/10.1016/j.tics.2006.12.002

Rauchbauer Birgit, Nazarian Bruno, Bourhis Morgane, Ochs Magalie, Prévot Laurent, & Chaminade Thierry. (2019). Brain activity during reciprocal social interaction investigated using conversational robots as control condition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1771), 20180033. https://doi.org/10.1098/rstb.2018.0033

Rogers, T. T., & Patterson, K. (2007). Object categorization: Reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General*, *136*(3), 451–469. https://doi.org/10.1037/0096-3445.136.3.451

Sebanz, N, Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76. https://doi.org/10.1016/j.tics.2005.12.009

Sebanz, Natalie, & Knoblich, G. (2009). Prediction in joint action: What, when, and where. *Topics in Cognitive Science*, *1*(2), 353–367. https://doi.org/10.1111/j.1756-8765.2009.01024.x

Shockley, K., Baker, A. A., Richardson, M. J., & Fowler, C. A. (2007). Articulatory constraints on interpersonal postural coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(1), 201–208. https://doi.org/10.1037/0096-1523.33.1.201

Skantze, G. (2016). Real-Time Coordination in Human-Robot Interaction Using Face and Voice. *AI Magazine*, *37*(4), 19–31. https://doi.org/10.1609/aimag.v37i4.2686

Strijkers, K., Costa, A., & Thierry, G. (2010). Tracking lexical access in speech production: Electrophysiological correlates of word frequency and cognate effects. *Cerebral Cortex (New York, N.Y.: 1991)*, *20*(4), 912–928. https://doi.org/10.1093/cercor/bhp153

Woollams, A. M. (2012). Apples are not the only fruit: The effects of concept typicality on semantic representation in the anterior temporal lobe. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00085