Contrastiveness in the context of action demonstration: an eye-tracking study on its effects on action perception and action recall

Amit Singh (amit.singh@upb.de)
Department of Psycholinguistics, University of Paderborn
33098 Paderborn, Germany

Katharina J. Rohlfing (katharina.rohlfing@upb.de)
Department of Psycholinguistics, University of Paderborn
33098 Paderborn, Germany

Abstract

The study investigates two different ways of guiding the addressee of an explanation - an explainee, through action demonstration: contrastive and non-contrastive. Their effect was tested on attention to specific action elements (goal) as well as on event memory. In an eye-tracking experiment, participants were shown different motion videos that were either contrastive or non-contrastive with respect to the segments of movement presentation. Given that everyday action demonstration is often multimodal, the stimuli were created with respect to their visual and verbal presentation. For visual presentation, a video combined two movements in a contrastive (e.g., Up-motion following a Down-motion) or non-contrastive way (e.g., two Up-motions following each other). For verbal presentation, each video was combined with a sequence of instruction descriptions in the form of negative (i.e., contrastive) or assertive (i.e., non-contrastive) guidance. It was found that a) attention to the event goal increased for this condition in the later time window, and b) participants' recall of the event was facilitated when a visually contrastive motion was combined with a verbal contrast.

Keywords: Attention; negation; contrastive guidance; eye-movements; action understanding; event representation

Introduction

In a given context, the conceptual structuring of an event depends upon what has already been provided in the past (Zacks & Swallow, 2007). While the upcoming sub-event within an event can be predicted through its history, the consolidation of the entire event at the end partly depends upon how the event segments relate to each other. One of the naturally occurring phenomena during communication in visual and verbal domains is the contrast between the constructions. This is realised when an expressed constituent of an event is characterised against the opposite of the same event. And if the two continuous sub-events are temporally bound by a status of occurrence and non-occurrence of a single instance, the prior event might act as the context, in light of which the later event would be interpreted. Interestingly, this property of event processing exists both in visual and verbal modalities. In the visual modality, the contrast can be realised using gestures of upward and downward motion, as if one is the opposite of the other (Hespos, Saylor, & Grossman, 2010). In verbal modality, negation can be used to create and infer, for example, the factual case against the non-factual (Kaup, Zwaan, & Lüdtke, 2007). This ubiquitous property within communication might not only allude to the nature of general cognition but also the context generating function of such elements.

In this study, we assumed that in natural settings, action demonstration involves both visual and verbal modalities, which can provide complementary information to facilitate event processing and memory. Although such multimodal communication is common in everyday life, there has been little research on how the combination of visual and verbal contrasts affects online event processing and memory. Therefore, our study aimed to fill this gap by testing the effects of multimodal contrasts on context creation and event memory. We based our rationale on previous work by Rolf, Hanheide, and Rohlfing (2009), who highlighted the importance of combining visual and verbal modalities to enhance the effectiveness of action demonstration.

To investigate it in the context of action observation, it is important to create stimuli in such a way that they simulate a task where two segments are overtly expressing the contrast. For this, we followed Hespos et al. (2010), creating perceivable transition between events for children by including rapid movements and goal knowledge. For the movements, we used spatial direction as one of the very salient properties where one could visually and verbally simulate the opposites easily. With respect to goal knowledge, it has been shown that the 14-months-old already segment action flow into subactions by looking proactively at the goal of each sub-task (Baldwin, Baird, Saylor, & Clark, 2001).

Linguistic negation has a similar property of generating a rich pragmatic effect by conceptually simulating a pair of representations, one of which the factual and the other alternate (Kaup, Lüdtke, & Zwaan, 2006; Kaup et al., 2007; Tian, Ferguson, & Breheny, 2016), such that when the latter is opposite of the former, both constructions make the same statement in an assertive–negative construction, for example, in Direction: Up–Down can have a counterpart of Up–Not Up. While at the surface level, both constructions seem to convey the same message, they can have a different repercussion for the attentional system since the later brings both factual and alternate in the focus of attention as shown by the previous studies (Kaup et al., 2006, 2007). Needless to say, this effect might be more prominent in a context of action demonstration where it is more plausible to negate an expression (Wason, 1965; Albu, Tsaregorodtseva, & Kaup, 2021), for example, when negation is used to convey a non-occurrence after an already occurred case, as opposed to a case where negation is used out of the context which generally has been

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

reported to elicit a negation processing cost (Clark & Chase, 1972; Carpenter & Just, 1975). Our study also built upon the findings in explanation literature, suggesting that contrastive explanations promote fine grained understanding and reduce cognitive load which might be a result of a rich context created by a conceptually contrastive information (Lipton, 1990; Miller, 2021).

Given that the language instructions during action demonstration have been shown to structure the action sequences, thereby influencing the action perception and learning (Rohlfing, Fritsch, Wrede, & Jungmann, 2006; Wrede, Schillingmann, & Rohlfing, 2013; Sciutti, Lohan, Koch, Gredebäck, & Rohlfing, 2016), our aim was to test the effect of contrastive guidance on the event memory in an action recall task.

Present Study

In present study, we used motion sequences of two kinds, a) visually contrastive, in which the occurrence of an event (ball moving up) was followed by non-occurrence of the same event (ball moving down) and b) visually non-contrastive, in which the same event occurred twice (ball moving twice Up or twice Down). The paths of these motion events were then paired with verbal descriptions of four types. a) assertion–assertion b) assertion–negation c) negation-assertion d) negation–negation.

Predictions

For the visually contrastive motion, a contextual effect would be expected when the occurrence and non-occurrence of an event is instructed by a sequence of assertion and negation, indicating that something previously happened but not this time, for example, a visually Up–Down motion can be instructed using [Now Up! – Not Up!]. A simple assertive instruction for the same motion will not create a similar contextual effect e.g., [Now Up! – Now Down!]. Crucially, for the contextual processing of negation, the negation needs an assertion in first instance so that it could be interpreted in terms of the prior event. Hence, in this condition, we would expect a better event memory for assertive–negative instruction condition.

For visually non-contrastive motion, we predict that a use of negation either at initial or later position will be detrimental for the recall, for example, in a visually non-contrastive motion, Up–Up, the verbal instruction [Now Up! – Not Down!] or [Not Down! – Now Up!] would lead to a to a processing cost because here the negation is not interpreted in terms of the already happened event. In this condition, we would expect a better memory for simple assertive instructions i.e., assertion–assertion.

For the visual attention, studies in the event processing have suggested that the event goal receives the maximum attention among other constituents of the event (Zacks & Swallow, 2007). Hence, we consider the time dependent fixation pattern to the goal object of our primary interest. Eye-movements in a scene is highly constrained by the task demands where a high processing cost at the cognitive level leads to a lower fixation (Liu, Li, Yeh, & Chien, 2022). Since the negation is primarily reported to evoke a processing cost (Clark & Chase, 1972), we aim to test its effect on the event goal which maximally remains in the target of attention. We predict that when the processing cost for a particular visual and verbal combination is low, then the attention to the goal will be higher. Considering this, in our setup, a combination of visually contrastive motion and assertion–negation instruction will lead to a higher fixation on goal due to contextual facilitation than other verbal conditions. For visually non-contrastive motion, we expect a higher fixation on goal in assertive–assertive condition. For other verbal conditions we do not have predictions.

Method

The methods reported in this study are approved by the Review Board of the associated university and the informed consent from all the participants were obtained prior to data collection.

Participants

Participants were 35 university students (Mean age = 23.90, *SD* = 2.97) years, recruited through advertisement in the classrooms and flyer distribution. All participants had native to fluent German proficiency and received 10 Euro for participation in the study. Data from 3 participants were discarded due to track-loss ($N = 2$) and failing to properly follow the experiment instructions ($N = 1$).

The sample size was based on previous eye-tracking studies on event understanding (Papafragou, Hulbert, & Trueswell, 2008; Bunger, Skordos, Trueswell, & Papafragou, 2016).

Stimuli

The stimuli were drawn from Hespos, Saylor, and Grossman (2009) and consisted of two pairs (N = 4) of short videos in which a ball was moved by a female actor with respect to the three landmark objects i.e., origin, midpoint and a final goal. For each video, the path of the ball was defined in terms of the sequence of directions it followed in two subsequent motions. The motion was classified as, **a)** contrastive, if – after reaching the midpoint – the ball moved in the opposite direction, i.e., Up–Down or Down–Up (Figure 1 (a) and (c)) and **b)** non-contrastive if the ball followed the same direction after reaching the midpoint i.e., Up–Up or Down–Down (Figure 1 (b) and (d)).

The path of the video was combined with verbal instructions indicating the direction of the motion either in assertive or negative condition. The verbal instructions were always congruent to the direction of the motion, for example an upward motion could either be instructed by an Assertion [Towards Up!] or a Negation [Not Down!], and likewise a downward by an Assertion [Towards Down!] or a Negation [Not Up!].

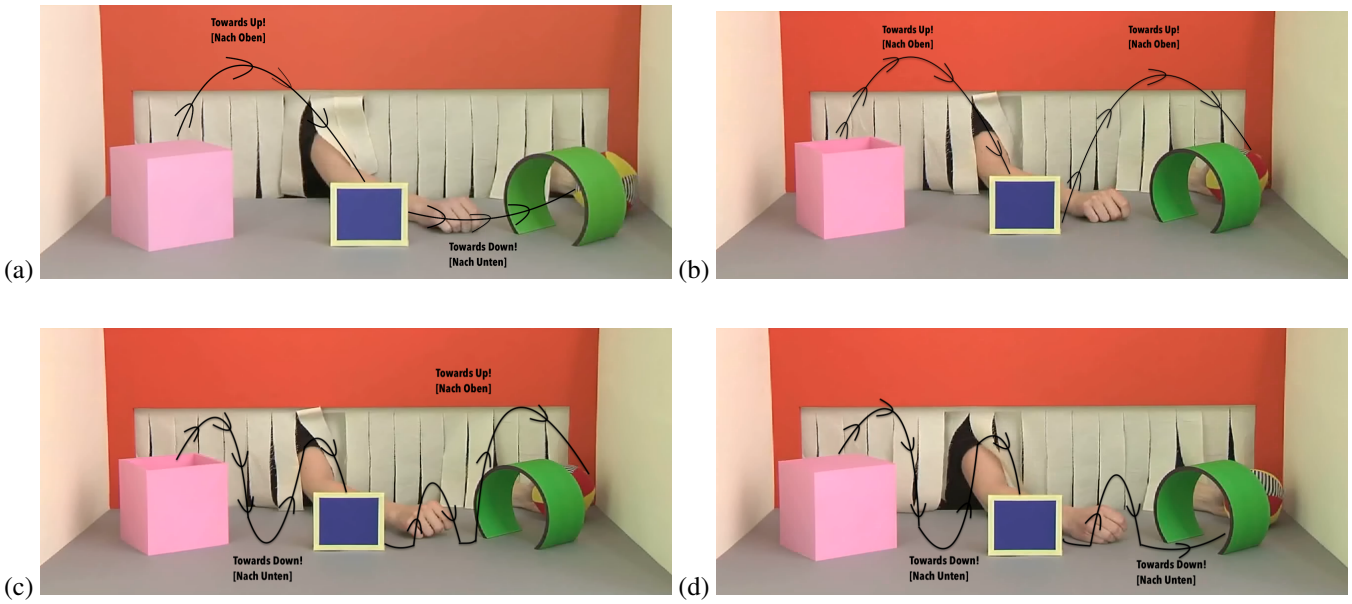

Figure 1: Stimulus videos depicting the path conditions; (a) **contrastive** [Up-Down], (b) **non-contrastive** [Up-Up], (c) **contrastive** [Down-Up], (d) **non-contrastive** [Down-Down]. Black lines show the movement path of the ball with instruction in **assertion-assertion** condition

The female researcher recorded verbal instructions using a tone of voice that is typically used when guiding someone through a series of movements. To add the verbal contrast on contrastive and non-contrastive motion types, the combination of verbal guidance were synchronised as, **assertion-assertion**, **negation-negation**, **assertion-negation** and **negation-assertion** where a video without instruction **(no voice)** was treated as a baseline condition. A total of twenty trials were created by combining 4 videos to 5 types of instruction conditions.

Procedure

The study was designed on Tobii-Pro software with 120 Hz remote eye-tracker installed on the 1920 (width) pixels X 1080 (height) pixels monitor. The size of the stimulus was 1080 (width) pixels and 720 (height) pixels. Participants were seated on a rotatable chair approximately 60 cm from the screen. The physical objects used in the video were placed behind the participants on a small stage. A video camera was installed behind the stage to record the action performance. The experiment started with a fixation screen followed by the first video then a probe message screen at the end. The probe screen asked the participants to perform the action shown in the video. Participants then turned back from the screen, performed the action on the small stage being seated on the same chair and then returned back to the initial position and proceeded to the next trial by pressing the space bar on eye-tracker keyboard. The presentation of the videos was randomised and the videos were presented in self-paced manner where participants were given unlimited time to perform the task and then come back to the initial position and continue to the next trial. The position of the chair was marked on the ground to keep the distance consistent between the eye-tracking screen and the participant. Prior to the main experiment, participants were given 10 practice trials to familiarise themselves with the experiment setup. The objects and its directions used in the practice session were different from the main experiment to avoid the carry-over effect. After the practice session the main experiment began. All trials were presented in random order without time restriction for the performance task. The entire experiment lasted approximately 30 minutes and the eye-gaze and videos data were recorded throughout the experiment.

Analysis

For the recall coding, videos of the action demonstration and performance were exported together from the eye-tracker and video recorder respectively. For the fixation metrics a separate data-set was exported from the Tobii-Pro Lab software.

Preprocessing of eye movement data

A circular area of interest (AOI) was created for goal object. Previous studies in event understanding have suggested that the goal receives greater attention than any other entity within the action (Hespos et al., 2009), which is also conceptualised as the path of the motion (Bunger, Trueswell, & Papafragou, 2012; Bunger et al., 2016; Bunger, Skordos, Trueswell, & Papafragou, 2021; Papafragou et al., 2008), where the action final object is represented as the motion path. Following this literature, the path AOI was defined by the area surrounding the goal object (here: green ring). In order to ensure that the moving object and path AOIs do not coincide, the AOIs

were only active from the beginning of the trial till the time the moving object reached the goal. Fixations were filtered using the Tobii-Pro-IV inbuilt algorithm. For each video the total trial duration was divided into 200 ms bins following the previous research suggesting that it takes around 200 ms to launch a saccade (Saslow, 1967). To calculate the fixation proportion at each time bin, the number of fixations falling in an AOI within a bin was counted and divided by the total number of fixations falling into that bin using a R script (R Core Team, 2022). Participants with more than 25% of trackloss across all the trials and trials with more than 50% trackloss were excluded from the analysis.

Pre-processing of recall videos

For recall, we compared the video data from the recorder with the video exported from eye-tracker. Each video was segmented into sub-actions based on the change in the path of the motion, which was ascertained by counting the inflection points in a path. The correct sub-action was scored 1 and incorrect as 0. The total correct sub-actions were summed at the end and then normalised over the total possible score to estimate the proportion. To ensure the reliability of the recall coding, the coding process involved two student assistants who independently coded the data. This approach was taken to establish inter-rater reliability between the two coders (Kendall's coefficient of concordance, $W = 0.78$).

Results

Eye movements

We were interested to test the time dependent changes to the fixation towards the goal for each combination of visual-verbal trials. The eye-tracking data was analysed using Growth Curve Analysis method, GCA (Mirman, Dixon, & Magnuson, 2008; Mirman, 2014). To get our dependent variable, we followed the prior work in motion event (Papafragou et al., 2008) and used the log transformed odd ratio fixation proportion on the goal AOIs for each 200 ms time bin and observed the preference to the goal across time. To capture the variability in the eye-movement, a 2nd Order GCA linear mixed effects model was fitted because the raw data roughly followed a U-shaped curve. The sum coding was done to treat - no voice - condition as the baseline against which other instruction conditions were compared. The model fitting procedure followed a maximal approach (Barr, Levy, Scheepers, & Tily, 2013) and included random intercept and slope adjustments to the subjects since a more complex model was unable to converge. All the analysis was done using lme4 package (Bates, Mächler, Bolker, & Walker, 2015).

Aiming to capture the change of attention towards the goal object in overall course of the trial, we examined the interaction of verbal instructions with polynomial time terms, i.e, linear and quadratic. If there is a significant effect of time on attention towards the path, as indicated by an increase or decrease in fixation, we expect to observe an interaction between that particular time term and the verbal condition. On the other hand, the absence of such an interaction would suggest that fixation remained stable with respect to a baseline..

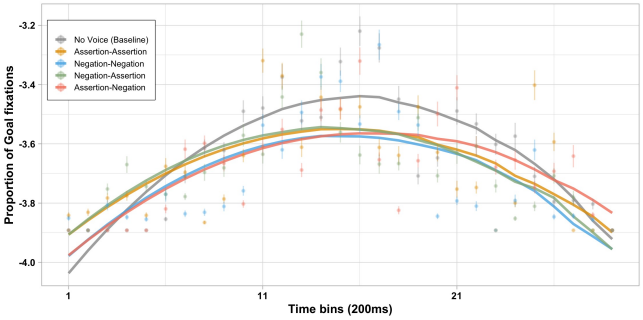

Figure 2: Proportion of Goal fixation for **Contrastive** motion

Table 1: Parameter estimates for the **contrastive** motion. The fixed effects interaction of instruction condition with linear and quadratic time terms

Fixed Effect	Estimate	*S.E.*	*t*	*p*
(Intercept)	-3.63	0.05	-76.55	0.00
Linear	0.20	0.03	6.98	0.00
Quadratic	-0.90	0.03	-31.99	0.00
Assertion-Assertion	-0.05	0.05	-0.87	0.39
Negation-Negation	-0.08	0.05	-1.58	0.12
Assertion-Negation	-0.06	0.05	-1.12	0.27
Negation-Assertion	-0.07	0.05	-1.25	0.21
Linear				
Assertion-Assertion	-0.16	0.04	-3.98	0.00
Negation-Negation	-0.15	0.04	-3.65	0.00
Assertion-Negation	0.02	0.04	0.46	0.64
Negation-Assertion	-0.34	0.04	-8.18	0.00
Quadratic				
Assertion-Assertion	0.31	0.04	7.82	0.00
Negation-Negation	0.23	0.04	5.81	0.00
Assertion-Negation	0.32	0.04	7.97	0.00
Negation-Assertion	0.22	0.04	5.49	0.00

For contrastive visual context (Figure 2), there was no main effect of verbal condition, suggesting that the overall time-independent looks towards the goal did not differ significantly across verbal conditions. To visualise the time dependent change, we analysed the interaction of linear and quadratic time terms with verbal conditions. There was a significant decrease in attention towards the goal when an instruction is provided alongside action (Table 1), which suggested that regardless of the characteristics of the verbal instruction i.e., negative or assertive, the attention towards the goal decreased with time when an action is accompanied by a verbal instruction. This is consistent with our prediction that the processing load leads to a decrease in attention towards the goal. Impor-

tantly, we found that this difference was not present in the assertion-negation condition ($\beta = 0.02$, SE = 0.04, t = 0.46) at the linear time term. This indicates that, for all verbal conditions, attention towards the goal object consistently decreased throughout the trial. However, in the assertion-negation condition, attention initially decreased but then increased at later times, and this pattern was not significantly different from the baseline (no-voice) at the linear time term. This finding might be interpreted as a reduced cost of negation processing following a positive context, which was comparable to the baseline (no voice) condition.

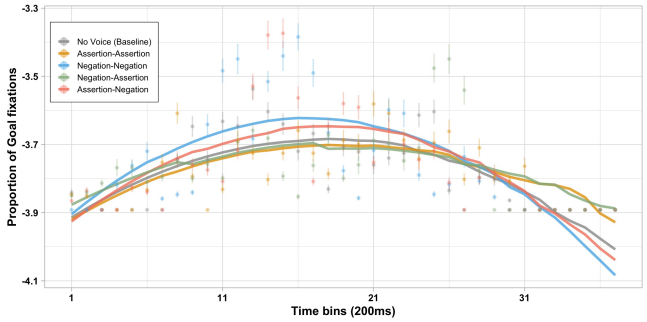

Figure 3: Proportion of Goal fixation for **non-contrastive** motion

Table 2: Parameter estimates for the **non-contrastive** motion

Fixed Effect	Estimate	*S.E.*	*t*	*p*
(Intercept)	-3.78	0.03	-108.23	0.00
Linear	-0.16	0.03	-6.10	0.00
Quadratic	-0.51	0.02	-20.68	0.00
Assertion-Assertion	0.01	0.04	0.19	0.85
Negation-Negation	0.03	0.04	0.78	0.43
Assertion-Negation	0.02	0.04	0.56	0.57
Negation-Assertion	0.04	0.04	0.92	0.36
Linear				
Assertion-Assertion	0.13	0.04	3.37	0.00
Negation-Negation	-0.15	0.04	-4.01	0.00
Assertion-Negation	0.00	0.04	-0.04	0.97
Negation-Assertion	0.17	0.04	4.63	0.00
Quadratic				
Assertion-Assertion	0.06	0.04	1.77	0.08
Negation-Negation	-0.19	0.04	-5.40	0.00
Assertion-Negation	-0.10	0.04	-2.98	0.00
Negation-Assertion	0.14	0.04	3.92	0.00

For the non-contrastive visual context, we did not observe a main effect of instruction (Figure 3), indicating that there was no significant difference in the time-independent fixation towards the goal when compared to the baseline condition. However, we did observe a significant interaction between the quadratic time term and all three verbal conditions, except for the assertion-assertion condition (Table 2). This indicates that providing an assertive-assertive instruction along with the non-contrastive action resulted in a similar fixation pattern towards the goal object as in the baseline (no-voice) condition. This is in contrast to the other verbal conditions where negation was used as an instruction, suggesting that there may be a processing cost associated with negation in a non-contrastive context. Additionally, we observed a similar pattern for the linear time term, with the exception of the assertion-negation instruction, suggesting that the processing cost of negation may be mitigated when it is preceded by an assertion.

Recall

Recall data was analysed using logistic regression model since the data followed a binary 0 or 1. The number of successes were calculated by summing all the correct responses, and the failure by subtracting the successes from the total score for each trial. The model was fitted using glmer function in R with binomial family, treating the fixed effects of instruction, path type and their interaction, and subjects as a random effect with varying intercept since a complex model with varying slope failed to converge. The instruction condition was sum coded to treat no voice condition as the baseline against which other conditions were compared. Similarly for the path type non-contrastive path was treated as the baseline. Results are provided below for path and instruction combinations (Figure 4).

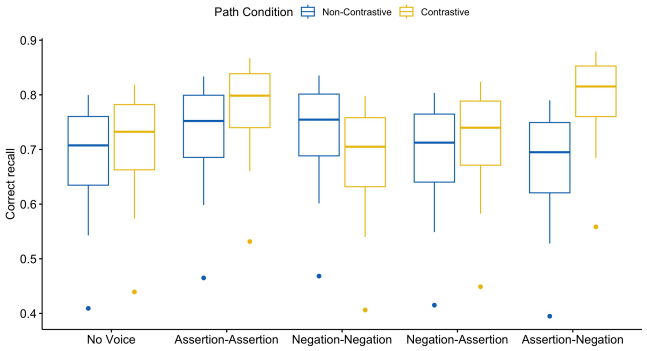

Figure 4: Proportion of correct recall for contrastive and non-contrastive path conditions for instruction conditions plotted against x-axis

The model showed no main effect for the instruction suggesting that participants' recall did not differ significantly for different verbal conditions (Table 3). There was no main effect of path condition as well suggesting that participants did not differ in their overall recall score across contrastive and non-contrastive paths. However, there was an interaction between path and verbal condition such that participants performed significantly better on contrastive path in presence of assertion–negation instruction ($\beta = 0.54$, SE = 0.23, t =

2.35, $p < 0.01$). A pairwise comparison within contrastive condition (Table 4) suggested a significant increase in recall for both assertion–negation and assertion–assertion condition; when compared with no verbal condition (baseline), this increase in performance was higher for assertion–negation condition (β = 0.47, SE = 0.16, t = 2.83, $p < 0.00$) than assertion–assertion condition (β = 0.37, SE = 0.16, t = 2.23, $p < 0.00$), however this difference was not significant.

Table 3: Parameter estimates of the fixed effects with instruction and path conditions as the predictor and recall as the outcome variable. Significant values are boldfaced

Fixed Effect	Estimate	*S.E.*	*t*	*p*
(Intercept)	0.83	0.14	5.97	0.00
Assertion-Assertion	0.23	0.16	1.43	0.15
Negation-Negation	0.24	0.16	1.51	0.13
Negation-Assertion	0.02	0.15	0.16	0.88
Assertion-Negation	-0.06	0.15	-0.39	0.70
Path: Contrastive	0.12	0.15	0.79	0.43
Assertion-Assertion	0.14	0.23	0.62	0.53
Negation-Negation	-0.38	0.22	-1.68	0.09
Negation-Assertion	0.01	0.22	0.06	0.95
Assertion-Negation	0.54	0.23	2.35	**0.01**

Table 4: Pairwise comparison for **contrastive** path with no verbal as baseline. Significant values are boldfaced

Fixed Effect	Estimate	*S.E.*	*t*	*p*
(Intercept)	0.96	0.14	6.70	0.00
Assertion-Assertion	0.37	0.16	2.23	**0.02**
Negation-Negation	-0.13	0.15	-0.86	0.38
Negation-Assertion	0.03	0.15	0.24	0.81
Assertion-Negation	0.47	0.16	2.83	**0.00**

For non-contrastive motion a similar pairwise comparison between instruction types treating no voice as the baseline showed no significant increase in performance for any instruction types (Table 5).

Table 5: Pairwise comparison for **non-contrastive** path with no verbal as baseline

Fixed Effect	Estimate	*S.E.*	*t*	*p*
(Intercept)	0.83	0.13	6.04	0.00
Assertion-Assertion	0.22	0.15	1.42	0.15
Negation-Negation	0.23	0.15	1.50	0.13
Negation-Assertion	0.02	0.15	0.15	0.87
Assertion-Negation	-0.05	0.15	-0.38	0.69

Discussion

This study aimed to explore the effect of contrastiveness in both visual and verbal modalities on action processing and event memory in comparison to a non-contrastive presentation. Our findings suggest that visual contrast alone does not enhance event memory. However, when a visually contrastive action is accompanied by a contrastive verbal guidance in the form of assertive-negative descriptions, it leads to better recall of the event. This implies that contrastive verbal cues can facilitate event memory in such cases. This finding is important for explanations that are usually multimodal.

We also tested whether there persists a distinct visual attention for event goal in this condition and found that the distribution of attention over the goal object decreases with time when the verbal instruction is provided alongside action — an effect similar to Sciutti et al. (2016), who found that the verbal instructions reduced predictive shifts towards the goal of action. The evidence also supports the findings of Papafragou et al. (2008), where it was reported that linguistic processing reduced the fixation towards the path (goal object) in relation to the manner. However, the current study remains limited with respect to the findings about the manner of the motion, where the manner could be conceptualised as the way the object traverses the entire trajectory. In this context, the movements of the hands can be categorised as the iconic gestures, indicating the relative position, which is specifically shown beneficial for communication (Holler & Beattie, 2003).

With regard to linguistic negation, the current study suggested that negation carries a processing cost, but its impact can be mitigated in a contrastive context where the prior information has been given. Our findings support the contextual effect reported by Albu et al. (2021) and expand it to the action demonstration scenario. Specifically, when a negative instruction follow an assertive instruction in a contrastive context, such as in [Now Up! - Not Up!], the negation can be interpreted in the light of the preceding event, leading to a more coherent sequence of events and improving recall. In cases where a later event is not related to a prior event, the use of negation does not improve recall. This is particularly evident in non-contrastive contexts where negation is processed without any contextual information. In contrast, assertive instructions seem to have a positive impact on event recall in such situations, likely due to the repetition of prior events.

Conclusions

This study sheds new light on a fundamental aspect of visual and verbal cognition - contrast. Our findings suggest that the combination of assertion and negation creates a rich contextual effect that can potentially reduce processing costs and enhance event memory for contrastive events. The findings highlight the importance of contrastive verbal guidance in facilitating learning of contrastive actions and offer possibilities for future research in event representation and action understanding.

Acknowledgements

This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824. We also thank four anonymous reviewers for their helpful suggestions.

References

Albu, E., Tsaregorodtseva, O., & Kaup, B. (2021). Contrary to expectations: Does context influence the processing cost associated with negation? *J Psycholinguist*, *50*, 1215–1242.

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development*, *72*, 708–717.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Keep it maximal: A case for using mixed effects models for language processing experiments. *PLoS One*, *8*(11), e79484.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bunger, A., Skordos, D., Trueswell, J., & Papafragou, A. (2016). How children and adults encode causative events cross-linguistically: Implications for language production and attention. *Language, Cognition and Neuroscience*, *31*, 1015-1037.

Bunger, A., Skordos, D., Trueswell, J., & Papafragou, A. (2021). How children attend to events before speaking: Crosslinguistic evidence from the motion domain. *Glossa: A Journal of General Linguistics*, *6*, 1-22.

Bunger, A., Trueswell, J., & Papafragou, A. (2012). The relation between event apprehension and utterance formulation in children: Evidence from linguistic omissions. *Cognition*, *122*, 135-149.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, *82*, 45–73.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472–517.

Hespos, S. J., Saylor, M. M., & Grossman, S. R. (2009). Infants' ability to parse continuous actions. *Developmental psychology*, *45*(2).

Hespos, S. J., Saylor, M. M., & Grossman, S. R. (2010). Infants' ability to parse continuous actions: further evidence. *Neural Networks*, *23*(8-9), 1026-32.

Holler, J., & Beattie, G. (2003). How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica*, *146*, 81-116. doi: 10.1515/semi.2003.083

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, *38*, 1033–1050.

Kaup, B., Zwaan, R. A., & Lüdtke, J. (2007). The experiential view of language comprehension: How is negation represented? In F. Schmalhofer & C. A. Perfetti (Eds.), *Higher level language processes in the brain: Inference and comprehension processes* (p. 255–288). Lawrence Erlbaum Associates Publishers.

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, *27*, 247-266.

Liu, J. C., Li, K. A., Yeh, S. L., & Chien, S. Y. (2022). Assessing perceptual load and cognitive load by fixation-related information of eye movements. *Sensors (Basel)*, *3*.

Miller, T. (2021). Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, *36*, E14.

Mirman, D. (2014). *Growth curve analysis and visualization using r*. Chapman and Hall / CRC.

Mirman, D., Dixon, J., & Magnuson, J. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475-494.

Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? evidence from eye movements. *Cognition*, *108*, 155-184.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rohlfing, K. J., Fritsch, J., Wrede, B., & Jungmann, T. (2006). How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, *20*, 1183-1199.

Rolf, M., Hanheide, M., & Rohlfing, K. J. (2009). Attention via synchrony: Making use of multimodal cues in social learning. *IEEE Transactions on Autonomous Mental Development*, *1*(1), 55-67.

Saslow, M. G. (1967). Latency of saccadic eye movement. *Journal of the Optical Society of America*, *57*(8), 1030–1033.

Sciutti, A., Lohan, K. S., Koch, B., Gredebäck, G., & Rohlfing, K. J. (2016). Language meddles with infants' processing of observed actions. *Frontiers in Robotics and AI*, *46*(3).

Tian, Y., Ferguson, H., & Breheny, R. (2016). Processing negation without context – why and when we represent the positive argument. *Language, Cognition and Neuroscience*, *31*, 683-698.

Wason, P. C. (1965). The contexts of plausible denial. *Journal of Verbal Learning Verbal Behavior*, *4*, 7–11.

Wrede, B., Schillingmann, L., & Rohlfing, K. (2013). Making use of multi-modal synchrony: A model of acoustic packaging to tie words to actions. In L. Gogate & G. Hollich (Eds.), *Theoretical and computational models of word learning: Trends in psychology and artificial intelligence* (p. 224-240). Lawrence Erlbaum Associates Publishers.

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, *16*, 80 – 84.