



Psycholinguistics
Faculty of Arts and Humanities
Warburger Straße 100, 33102 Paderborn, Germany

From Negation to Contrast and Context: A Framework for Contrastive Verbal Guidance of Actions in Dynamic Contexts

Dissertation for the award of the academic degree
Doctor of Philosophy (Dr. phil.)

in the subject of Psycholinguistics
at Paderborn University

by

AMIT SINGH

Supervisor

PROF. DR. KATHARINA J. ROHLFING
Paderborn University

Second Examiner

PROF. DR. BARBARA KAUP
University of Tübingen

Examination Committee

PROF. DR. HEIKE M. BUHL
DR. ANDREA KARSTEN

© 2026 by the Author(s)

Permission is granted to make digital or hard copies of this work for personal or classroom use free of charge, provided that copies are not made or distributed for profit and that they include this notice and full citation. Any other use requires prior written permission of the author.

Doctor of Philosophy (Dr. phil.) by Amit Singh

Supervisor:	Prof. Dr. Katharina Rohlfing
Second examiner:	Prof. Dr. Barbara Kaup
Examination Committee:	Prof. Dr. Heike M. Buhl Dr. Andrea Karsten

Paderborn, January 9, 2026

Eidesstattliche Erklärung

Hiermit erkläre ich gemäß §12 der Promotionsordnung:

- a) dass die vorgelegte Arbeit selbstständig und ohne Benutzung anderer als der in der Arbeit angegebenen Hilfsmittel angefertigt wurde;
- b) dass die Arbeit bisher weder im In- noch Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt wurde;
- c) dass früher oder gleichzeitig kein Promotionsverfahren bei einer anderen Hochschule oder bei einer anderen Fakultät beantragt wurde, gegebenenfalls nebst vollständigen Angaben über dessen Ausgang.

Paderborn, July 6, 2026



Amit Singh

Preface

From Negation to Contrast and Context, the title outlines the story of my investigation timeline. I started as a psycholinguist studying negations to investigate contrastive explanations, particularly to make them better suited for explainability (XAI) research. However, by going beyond purely linguistic approaches and through an interdisciplinary dive into multifarious aspects of contrast, the investigation turned out to be a more comprehensive theory that is not only constrained to linguistics, such as negations, but in general tries to characterize its dynamics along non-linguistic contexts. It seems that contrast can better be understood as a form of context, similar to other forms of contexts. This should by no means be taken as something special or detached from other forms of contexts. In this thesis, I attempt to organize the idea of contrastiveness on different levels, as we can observe in our day-to-day perception. Starting from negation as a central linguistic device which can give rise to explicit contrast at the linguistic level, the idea was then to look into other ways of language framing that can also be perceived as contrast, and later, into domain-general contrasts at different non-linguistic dimensions.

The combined synthesis eventually led to a *contrastiveness scale*, motivated by the idea of relational perception, understood as how Ground stands for and against Figure within perception, drawing on influential ideas from Gestalt psychology. I argued that contrastive contexts are not merely “there” as a static background, but are dynamically developed and reshaped as people act, speak, and repair one another during interaction. On the empirical side, all the aforementioned theoretical and practical notions are derived from a series of studies conducted over the past four years to investigate contrast across multiple levels (linguistic and non-linguistic), dimensions (spatial, eventive, and so on), and settings (interactive and non-interactive). These studies jointly trace how contrastive structures emerge, how they are used, and how they can be foregrounded or backgrounded.

On the application front, the thesis also lays the groundwork for modeling within a cognitive architecture based on contrastive guidance, known as ContACT-R, which is further developed into an interactive account, the InterACT-R framework. Here, contrast is treated as an organized resource for scaffolding that can be tailored to the learner’s capabilities and the available or emerging context. The model also seeks to connect contrastive guidance to action performance, for instance by tracking task success and failure over time and adapting guidance accordingly. Although concrete proof-of-concept realizations are left open as future work, my hope throughout this development has been to utilize findings across experiments to derive contrastive scaffolding strategies that are tailored to users’ capabilities, informed by their interaction history, and essentially grounded in cognition.

Acknowledgments

It would not be an exaggeration to say that some of the most pivotal moments of my academic life have unfolded while I was writing my thesis. This thesis is no exception. As I am writing, large language models are emerging as a new genre of AI, redefining knowledge production and learning. Similarly, my master’s thesis was also written during an unusually decisive period when COVID-19 prevailed and my plans for a PhD felt uncertain and blurred. In that crucial period, I sincerely want to thank my supervisor, Prof. Dr. Katharina Rohlfing, for giving me the opportunity to work in the interdisciplinary and collaborative environment of TRR318 “Constructing Explainability”. Coming from Cognitive Science, joining an interdisciplinary center like TRR318 was one of the best opportunities I could have hoped for. I am especially grateful for the intellectual openness and the trust she placed in me, which allowed me to pursue project-oriented research while still developing my own ideas, ultimately helping me carve my own research pathway.

I want to gratefully thank Prof. Dr. Barbara Kaup for agreeing to provide feedback on the first draft of this thesis, which will certainly help me improve it as I go along.

I would also like to thank the members of the RTG TRR318, who supported me from the very beginning by helping me to navigate the initial formal academic requirements and offering training programs and workshops during my first days. This collaborative environment not only provided interdisciplinary perspectives, but also strengthened my understanding of research management practices that will remain invaluable for the future.

I would like to extend my thanks to my A05 project partners and all associated members. Without our discussions, many of the early ideas (naive and loosely connected) in the first year could not have developed into richer conceptualizations. I am also grateful to my colleagues at the SprachSpiellabor, who supported me through daily practical challenges during my transition and helped with data acquisition and logistics needed to run experiments. I thank them for numerous discussions that helped me shape new ideas, design better experiments, and approach the problem from alternative perspectives. Special thanks to Dr. Angela Grimminger, who supported me whenever I needed bureaucratic help and shared her knowledge, especially regarding data management skills. I also want to thank my amazing student assistants (Alana, Annabell, Tim, and Lasse). Their support was essential to this work and made the data collection possible.

This research was generously funded by the DFG (German Research Foundation), and I sincerely acknowledge their continued support throughout my research journey. Last but not least, I want to thank my parents and all my family members whose support has always been there for me throughout the entire journey, come rain or shine.

Contents

List of Figures	xi
------------------------	-----------

List of Tables	xvii
-----------------------	-------------

1 Introduction	1
1.1 Contrastive Contextualism	4
1.2 Dimensions	6
1.2.1 Situative Contrast	8
1.2.2 Functional Contrast	11
1.2.3 Locative Contrast	13
1.2.4 Eventive Contrast	15
1.2.5 Social Contrast	18
1.3 The multimodality of contrast	20
1.4 Roadmap and contributions	23
1.4.1 Publications and summary	23
2 Language as a contrastive mechanism	29
2.1 Does language foreground relations?	29
2.2 Linguistic contrasts steering attention	31
2.3 Implicit and explicit contrasts	38
2.4 Negation	40
2.4.1 Negation across contrastive dimensions	46
2.4.2 Formalization of contrastiveness scale	51
2.5 Contrastive Guidance	55
2.5.1 Design principles for contrastive guidance	58
2.5.2 Discussion	61
3 Context	65
3.1 Why Context?	65
3.2 Approaches to studying context	68
3.2.1 Context is what is available	69
3.2.2 Content based	70
3.3 Context is what is developed	78
3.3.1 Context based on dynamicity	80

3.4	Interactivity forming its context	85
3.4.1	Discussion: why context matters for contrastive guidance	89
4	Interaction and Scaffolding	93
4.1	Studying interaction	93
4.1.1	The traditionalists	95
4.1.2	The interactionists	96
4.2	HRI as a field to study language and visual processes	97
4.2.1	Language processes	99
4.2.2	Visual processes	101
4.3	Scaffolding	103
4.3.1	Faded scaffolding	104
4.3.2	An interactive model of scaffolding	107
4.3.3	Discussion	115
5	Conclusion and outlook	117
5.1	Limitations	119
5.2	Future directions	120
A	Relation to publications	123
	Bibliography	127

List of Figures

1.1	A contrastive context. The figure compares a normative situation (a), where a dustbin is placed on the sidewalk consistent with experience, against a contrastive situation (b), where the dustbin appears on the crossing. Spatial implausibility highlights the relational prominence over the object's intrinsic attributes.	3
(a)	Situation A	3
(b)	Situation B	3
1.2	Figure depicting continuum of contrastiveness. Deviations from normative contexts (right) to implausible contexts (left) increases relational prominence over figural attributes.	5
1.3	Dimensions of relations. A visualization of the five key domains where relational properties emerge: Situative (spatial configuration), Functional (learned affordances), Locative (exhibiting alternate locations), Eventive (spatio-temporal expectancies), and Social (social scripts).	7
1.4	Examples of situative contrast. The central node represents the Ground (situative expectation), while the surrounding images depict Figures (realized situations) at varying distances from this ground. The length of the arrow corresponds to the magnitude of contrastiveness or relational prominence.	10
1.5	Examples of functional contrast. The central node represents the Ground (functional expectation or affordance schema), while the surrounding images depict Figures (realized actions) at varying distances from this ground. The arrow length depicts the degree of contrast from the canonical function.	12
1.6	Examples of locative contrast. The central node represents the Ground (locative expectation or available set of placements), while the surrounding images depict Figures (realized placements) at varying distances from this ground. The arrow length depicts the degree of spatial competition or deviation from the canonical placement	14
1.7	Examples of eventive contrast. The central node represents the Ground (eventive expectation established by history, e.g., a repeating pattern), while the surrounding images depict Figures (realized next actions) at varying distances from this ground. The arrow length signifies the degree of deviation from the projected sequence	16

2.1	Stimuli depicting the action conditions. The blue dashed line indicates the trajectory of the ball relative to the objects. Panel (a) shows the Noncontrastive (OVER-OVER), where the action is repeated. Panel (b) shows the Contrastive (OVER-UNDER), where the second action creates a spatial contrast against the first. Adapted from Singh and Rohlfing (2025a).	31
(a)	Noncontrastive (OVER-OVER)	31
(b)	Contrastive (OVER-UNDER)	31
2.2	Relationship between recall score and goal-attention over time. The heatmap illustrates the probability of looks towards the goal (indicated by the color gradient and contour lines) as a function of time (x-axis) and recall score (y-axis). Lighter regions with white shading indicate a higher probability of goal-looks, showing how attention to the goal during the event correlates with later recall performance. Adapted from Singh and Rohlfing (2025a). .	32
2.3	Two forms of multimodal synchrony (Singh & Rohlfing, 2025a): (A) sequential and (B) simultaneous. In the sequential presentation (Experiment 1), attention is directed toward the immediate action (path) as the instruction precedes and introduces the action. In the simultaneous presentation (Experiment 2), overlapping instruction and action are expected to shift attention directly to the action goal due to ever-present goal bias in event. .	34
2.4	Response times for questions about missing information. The bar chart displays the mean response time (in seconds) for questions targeting the manner of the action (orange bar) versus the path of the action (gray bar). The results indicate that path-related questions were answered significantly faster than manner-related questions. Adapted from Singh and Rohlfing (2024)	37
2.5	An example of negation being more impactful than assertions. The image shows a handwritten note placed above a switch, reading "Heizungs Notschalter! Kein Lichtschalter!!!" (Switch for the heater! Not for lights!!!). This illustrates how negation ("Kein..") is used to explicitly inhibit a prepotent response (turning on what looks like a light switch) by opposing the expected function with the actual one, often conveying stronger caution or emphasis than a simple assertion would.	41
2.6	Recall for contrastive and noncontrastive motion events as reported in Singh and Rohlfing (2025a)	44
(a)	Experiment 1	44
(b)	Experiment 2	44
2.7	Effect of multimodal synchrony on goal-attention as reported in Singh & Rohlfing 2025: Comparison of goal-looks over time for instructions provided before the action (Exp1) versus instructions provided along with the action (Exp2), across both Contrastive (top panel) and Non-contrastive (bottom panel) contexts.	46

2.8	Posterior distributions for contrastive instruction on cognitive load (pupil dilation) in the eventive contrast. Negative distribution (middle row) indicates participants' lower pupil dilation (reduced cognitive load) when processing Negation compared to Assertion after a repetitive action history. . .	50
2.9	Schematic ACT-R style model of contrastive guidance (ContACT-R). The current context C is encoded as dimension-specific chunks $F_d(C)$ (Figure). Declarative memory provides canonical templates G_d (Ground), which are retrieved as $G_d(C)$ for the current context. A comparison stage computes distances $\Delta_d(C)$ and relational prominence values $R_d(C) = \Delta_d(C)A_d(C)$, where $A_d(C)$ captures the availability of alternatives. These values together feed into a guidance policy that in a context chooses between implicit devices and explicit contrast with negation ($Neg_d(C)$).	54
2.10	Relational prominence $R_d(S)$ as memory-based distance on the contrastiveness scale, and its mapping to guidance choice. At low values, the scene is script-consistent and implicit devices suffice. As $R_d(S)$ increases, alternatives become live and explicit contrast with negation becomes increasingly beneficial, especially for highly contrastive, alternative-rich cases.	55
3.1	Schematic relation between context and verbal encoding. In the upper panel, a relatively sparse context (small circle) supports only a coarse verbal encoding (larger, more elaborated box): much of what is needed for understanding must be spelled out in language. In the lower panel, a rich context (large circle) allows for a compact utterance (smaller box), since many presupposed conditions are already provided by the situation and common ground.	67
3.2	Schematic of content-based context. The diagram illustrates how context encloses the Figure–Ground relation. Crucially, the content of context is not arbitrary but structured along specific dimensions (right panel: situative, functional, locative, eventive, and social), which determine which Ground is available for interpreting the Figure.	71
3.3	Experimental setup for investigating the role of eventive context in Human–Robot Interaction (adopted from (Singh et al., 2025)). The left panel shows a participant interacting with a robot while gaze is recorded. The right panel depicts the workspace in which a sequence of object placements (screws) establishes an action history, or Ground (e.g., a repetitive vertical manner), against which the target action is performed and interpreted. . . .	72

3.4	Visualization of the two types of eventive contexts used in the study (Singh et al., 2025). The top row depicts the Simple situational model, where the action history is uniform (all screws placed vertically), establishing a single, unambiguous ground. The bottom row depicts Rich situational model, where the history includes mixed manners (vertical and horizontal), thereby keeping multiple relevant alternatives available in the context before the critical trial.	74
3.5	Gaze dwell times on the context prior to clarification requests. The histograms display the distribution of dwell times on the physical context for the Simple (left, red) and Rich (right, green) situational models. Participants exhibited significantly longer dwell times in the rich condition, indicating that when the context held diverse information (mixed manners), participants utilized it as an external resource, engaging in extensive "back-referencing" to resolve the under-specified instruction (Singh et al., 2025). .	76
3.6	Repair types by context. Participants mostly used open repairs (e.g., "How?") in the Simple context. In the Rich context, where multiple alternatives were visible, restricted repairs (e.g., "Vertical or Horizontal?") significantly increased, showing that richer contexts support more precise requests. . . .	77
3.7	Temporal distribution of gaze looks relative to repair onset. The plots show looks to context areas over time for Simple (left) and Rich (right) models. The bottom-right panel highlights that Restricted Repairs in Rich contexts are preceded by dense visual scanning (back-referencing) of the available alternatives, whereas Open repairs (top row) show sparser attention to the context (Singh & Rohlfing, 2025c). The vertical blue line depicts the repair onset.	78
3.8	Adopted from Singh & Rohlfing, 2024. Looking behavior for two different partner models. Left plot: unsuccessful scaffolding; Right plot: successful scaffolding. Attention on the face and torso was more distinctive where the robot successfully provided a resolution to the repair initiated by the participant as opposed to where the response was absent. The shaded region shows the timeline of repair onset and offset.	82
3.9	A depiction of a Visual World Paradigm setup. Participants are seated in front of a screen wearing an eye-tracker, while stimuli are presented in a trial-by-trial fashion (different trials are shown on the right margin). Cognitive processes are hence measured in a static context as the trial resets.	83

3.10	The diagram maps different interactive settings based on the variability of their Task Model (x-axis) and Partner Model (y-axis). Traditional Visual World Paradigms (bottom-left) rely on static grounds with low dynamicity, whereas real-world Social Interactions (top-right) involve highly dynamic, co-constructed models. HRI studies occupy the intermediate space, offering a methodological advantage where one dimension can be kept relatively stable to act as a backbone, while the other is allowed to evolve dynamically.	86
3.11	The coupling of Partner and Task models. The plots illustrate how the perception of the partner (fixation on the face) influences task performance (question onset time). From Singh and Rohlfing (2024)	87
4.1	Temporal dynamics of social gaze during instruction (Singh & Rohlfing, 2024). The plots display the proportion of fixations on the Robot's Face (red) and Task Objects (teal) over time. (A) Full Information: A routine pattern where gaze shifts to the object to execute the task and returns to the face for confirmation. (B) Missing Information: When the robot's guidance is under-informative, participants rapidly shift their gaze back to the robot's face (approx. 4s), visually signaling trouble and bidding for attention before the onset of their verbal repair question (red dotted line).	102
4.2	Faded scaffolding adapted to human-robot interaction. The diagram illustrates the application of Zone of Proximal Development (ZPD) to an HRI context. The robot acts as the More Knowledgeable Other, providing high-level support (scaffolding) when the learner is within the ZPD. As the learner internalizes the task and moves toward the center (Independent Ability), the robot's guidance is systematically reduced ("faded") to promote autonomy.	104
4.3	Figure 4.3: Distribution of tutor support versus learner autonomy, illustrating the fading effect as the step in the task increases. The fading resets as the new task is introduced.	105
4.4	Illustrative relations between degree of scaffolding fading and learner understanding.	107
4.5	Logistic mapping from evidence minus threshold to the probability of contrast on dimension d .	113
4.6	InterACT-R as an interactive extension of the ContACT-R model. The left-to-right backbone (Context $\rightarrow$ Encoding $\rightarrow$ Contrastiveness) is preserved; an interaction state S_t and task feedback loop modulate the guidance policy over time.	114

List of Tables

- 2.1 Linguistic mechanisms modulating relational prominence. Few examples of how different constituents (e.g., verbs, negation, aspect) structure the Figure–Ground relation to shift perception along the contrastiveness scale. . 39

Chapter 1

Introduction

Our perception is relational. We cannot help but see things around us in relation to other objects or events, making the isolated perception nearly impossible. In this sense, the meaning of an entity is constitutive of its surroundings that are also partly experiential in nature (Barsalou, 1999; Gibson, 1979; Rohlfing et al., 2003; Varela et al., 1991; for a review Hafri & Firestone, 2021). The idea traces back to *Gestalt* principles where the mind ascribes global interpretation to an entity derived from holistic perception of emerging features in an environment (Koffka, 1935; Wertheimer, 1923). In the literature, such an emergence is usually tagged with different theoretical constructs like, contexts (Dourish, 2004; Duranti & Goodwin, 1992), situations (Barwise & Perry, 1983; Rohlfing et al., 2003), background (Talmy, 1975, 1985a), frame (Goffman, 1974; Minsky, 1975) and settings (Barker, 1968), but also call it relations when scope just spans around few objects or features in the vicinity (Gentner, 1983; Landau & Jackendoff, 1993). Given the conflated meanings of these emergences, I would use *context* as a generic term to refer to any relational experiences hereafter. The idea that an entity deriving meaning from its surrounding and activating experiences is intuitive and commonsense, however the starting point of most analyses is based on isolated objects, where the objects features and its attributes become the central part, an approach popularly known as symbolic account (Fodor, 1975, 1983; Newell, 1990; Newell & Simon, 1976).

Whereas relational perception in an environment is an outcome of experience, since it is often learned through our interactions (Barsalou, 1999; Gibson, 1979), attributes are purely symbolic constructs that merely characterize the entity of interest (Harnad, 1990; Newell, 1980). Similar representational distinctions have also prevailed over the last few decades in cognitive science, often framed as grounded vs. symbolic (Barsalou, 2008; Fodor, 1983), concrete vs. abstract (Barsalou et al., 2018), and, more recently, modal vs. amodal (Kaup et al., 2024). Nevertheless, given that, at a deeper level, relations are subject to arbitrariness—where intrinsic factors such as experiential traces, simulation, and predispositions complicate the matter (Gentner, 1983; Hofstadter, 1995)—intuitively we are easily predisposed to start with symbolic analysis, simply because it seems easier than a relational approach (Ockham, 1974). In such cases, the classical “divide and conquer” strategy best suits the contextual analysis (Marr, 1982). Seen from another perspective, the starting point of contextual analysis can also be at relational level, that is, by reasoning

on an entity’s very existence against a context rooted in experiential traces rather than on symbolic attributes (Bar, 2004; Fischer & Zwaan, 2008; Oliva & Torralba, 2007; for recent review Kaup et al., 2024). For example, when we encounter an object (e.g., a dustbin), our perception can begin with relational analysis, which is more grounded in our experiences than abstract symbolic attributes such as color, texture, or size (Pecher & Zwaan, 2005). This duality of representations may seem obvious, in the sense that taking both natures into account would yield a more complete picture (sometimes referred to as a hybrid view (Kaup et al., 2025)). Yet the issue here is one of priority and order, namely whether relational perception follows symbolic analysis or whether symbolic description dominates across most contexts.

In this thesis, I argue that whether one form dominates over the other depends on what the context has to offer and on what our experiences bring into contextual analysis. More specifically, I suggest that, at one level, almost all contextual interpretation of an entity can be derived through a particular form of relational identity, called *contrast*, essentially having its origin in experiential traces (Murphy, 2002; Tversky, 1977). Here we shall ask why contrast? Contrasts seem to be foundational for relational perception and thereby its interpretation, especially due to its unique property of exhibiting the entity’s attributes of concern (called figure) but also richly embodying information from its experiences (called ground) providing a rationale for figure’s existence against specific ground.¹ This is not only true for visual but also for linguistic perception (Langacker, 1987; Talmy, 1985a). Relational perception thus begins by considering the reasons an entity is present in the environment at all. The very existence of an entity in the environment comes with a reason, which is often teleological in the sense that at the core it serves a relational purpose (Austin & Vancouver, 1996; Dennett, 1987; Kelemen, 1999). Imagine a context such as in Figure 1.1; if asked which of the contexts out of the two are best suited for representing contrast, most answers would align with Figure 1.1b. Crucially in both the figures, the principal entity (dustbin) acts as the figure and the rest of the scene is perceived as background. However, the prominence of the background varies depending upon how the figure is construed in the given context. Meaning, an entity observed against an experientially implausible background does not only exhibit properties of the entity itself (attributes related to the dustbin) but at the same time also makes the background more prominent (why is the dustbin lying on the crossing?), which otherwise would not become salient in normative situations (Biederman et al., 1982; Loftus & Mackworth, 1978). This simple example elucidates how we take our everyday experiential traces for granted in most contexts and might not reason for all the constituents and their attributes available in a context, unless some form of contrastiveness is perceived, making the relational traces prominent over attributes. One of the earliest accounts of how perception organizes object

¹For literary consistency, the *Figure–Ground* distinction is closely related to Gestalt psychology (Rubin, 1915) and was later extended in the seminal work of Talmy (1975, 1985a). In particular, whereas Talmy (1975, 1985a) treats the Ground largely as a physical property of space, in this thesis I define it as a product of experience which are essentially abstract without specific geometric representation.

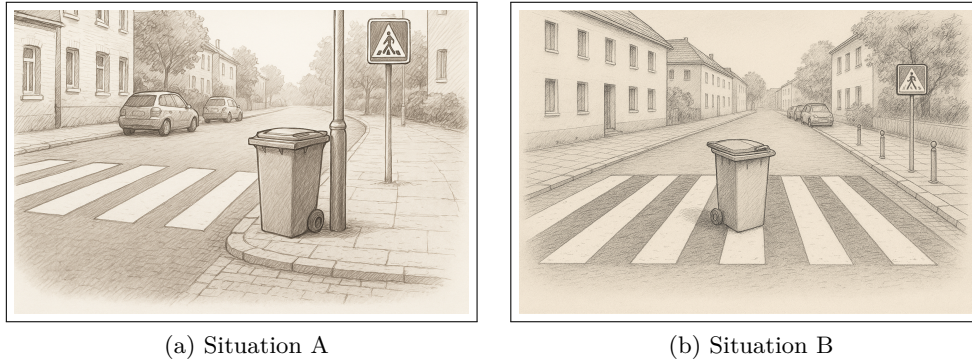


Figure 1.1: A contrastive context. The figure compares a normative situation (a), where a dustbin is placed on the sidewalk consistent with experience, against a contrastive situation (b), where the dustbin appears on the crossing. Spatial implausibility highlights the relational prominence over the object’s intrinsic attributes.

attributes and object relations comes from the structure-mapping framework of Gentner (1983). It argues that relations, rather than attributes, provide the primary basis for interpretable mappings from a base to a target domain, thereby going beyond purely feature-based interpretations (Tversky, 1977).

Hence, although relational and attributional elements are ever-present in almost any context, we might not bring both of them into active analysis to the same degree unless the dynamics between them become either complementary or disruptive as in Figure 1.1a. The choice is not deliberate but ecological, provided that in most situations our attention is also shown to be selective (Desimone & Duncan, 1995; Driver, 2001). Thus, considering a continuum of contrasts, most subtle contrastive perceptions might not be realized in everyday situations, although contrastive perception is a way for holistic analysis of a context by serving twofold functions: foregrounding the relational aspects and highlighting the attributional elements simultaneously (Garner, 1974; Minsky, 1975). Whereas for attributes, contrastive relations are already identified through physical properties like colours and shapes (Itti et al., 1998; Tversky, 1977), semantically such relations can be defined on multiple dimensions depending on where they lie on the continuum scale and how they are brought into attention through the relational and attributional dynamics acting for and against each other (Gentner, 1983; Goldstone, 1994; Talmy, 2000). For example, when an entity is observed at a semantically and experientially implausible situation, it not only reveals about the attributes of the entity but also makes the experiential and relational components more prominent (Biederman et al., 1982; Henderson & Hollingworth, 1999). These relations indeed cohere to semantics of the scene whose properties can be defined on some dimensions. On this view, *contrast* can be defined as an explicit relational operation between two (or more) alternatives where discourse entities or propositions are juxtaposed as comparable members of a contextually relevant alternative set, making their differences

salient to the addressee (Repp, 2010; Tversky, 1977).

In this thesis, my aim is to first provide generic dimensions in everyday cognition (non-exhaustive at this point) at which experiential relations act to exhibit contrastive relations on a continuum and then attempt to explore how language might complement these contrastive perceptions (Gentner, 2003; Gleitman, 1990; Talmy, 1985a; Wolff & Holmes, 2011). Supplemented by experimental results, we also identify *negation* being a central linguistic constituent eliciting and complementing contrastiveness (Horn, 1989; Just & Carpenter, 1971; Kaup & Dudschig, 2020a). Provided multifarious definitions of contexts characterization and at the same time, very few methods to formalize its properties (Dourish, 2004; McCarthy, 1993), in this thesis, I aim to provide a theoretical pathway to analyze the contexts through relations and validate their perceptions on given dimensions. The overarching goal of the project was to develop linguistic scaffolding strategies to guide contrastive perception for learning and situational understanding (Bruner, 1978; Vygotsky, 1978). Motivated by this goal, I will present two representative studies at the later part of the thesis looking along this development. Linguistic negation being the crucial part of contrastive guidance deserves a dedicated section and deals especially with its rich pragmatic effects reported in the literature (Kaup & Dudschig, 2020a; Wason, 1965).

1.1 Contrastive Contextualism

All contexts are contrastive, but some contexts are more contrastive than others.

— after Orwell

Contrast is ubiquitous, but its salience varies depending on how relational dynamics, building upon experiential and simulational traces, make themselves prominent over isolated symbolic and attributional elements (Gibson, 1979; Itti et al., 1998; Minsky, 1975). At any spatiotemporal dimension, an entity’s state of affairs is defined by its *attributes* and *experiential traces* drawn from past sensory experiences (Barsalou, 2008). For example, we might define Figure (1.1, a) by listing some attributes (e.g., GREEN(dustbin), car, road, pedestrian, zebra crossing) and some relations (e.g., ON(dustbin, pedestrian), or dustbin *kept on* the pedestrian crossing *near* the pole). Similarly, if we compare this description with Figure (1.1, b), one might be disposed to first highlight the relational constituent to seek a reason for the dustbin’s existence on the crossing, even before the isolated attributes. Why does it happen that one context seems to highlight relations more than the other?

While walking on the road, we might never bother about the dustbin placed on the side of the road unless it becomes spatially (such as implausible positioning) or temporally prominent (new introduction) with respect to our experiential background. Pragmatically, this can translate to a state of affairs on the spatial dimension such as: “How come the dustbin is on the crossing?” and for the temporal as: “There seems to be a new dustbin today which was not there before”. Spatial implausibility can also preclude the temporal

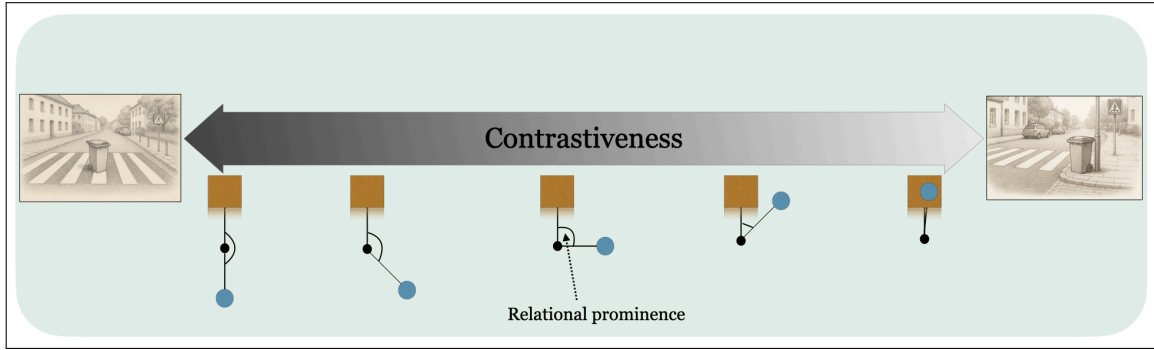


Figure 1.2: Figure depicting continuum of contrastiveness. Deviations from normative contexts (right) to implausible contexts (left) increases relational prominence over figural attributes.

component, such as: “The dustbin used to be on the roadside but today it is in the middle.” When relational dynamics between figure and ground become more relevant for the situation than the mere figure (or its attributes) itself, experiential traces give rise to spatiotemporal contrastiveness (Garner, 1974; Talmy, 2000).

Though in normative situations, we might not see these relations becoming as prominent as for experientially implausible situations, they nevertheless are an inherent part of the context. Hence, the relational prominence between the principal entity (such as figure) and environmental experiences (such as ground) establishes a degree of contrast on a continuum of *contrastiveness* that is sensitive to how the figure and its experiential grounding cohere at some spatiotemporal dimension (Figure 1.2). As illustrated in Figure 1.2, on one end lies the normative situation; on the other end, implausibility increases the relational prominence over the entity itself (Loftus & Mackworth, 1978). In situations with high relational prominence, people are more likely to highlight the relations (dustbin on crossing) than the entity or its attributes (green dustbin) (Clarke et al., 2013; Viethen & Dale, 2008). Phenomenologically, context perception is subjective, with individuals evoking different experiences in a situation; however, at the two extremes (highly normative or implausible), perception shall exhibit uniformity across the population. For example, situations at the extreme left (implausibility) would often evoke the relational component more than the extreme right (normative contexts). This is in line with a wide range of studies suggesting that the principal entity or figure stays at the prominence and the background exists there to support its interpretation and more often does not come into the locus of attention (Mazza et al., 2009; Rubin, 1915). I shall here highlight the importance of contrastiveness as anchoring the relational prominence over the figures.

In line with this research, I would present studies supporting the effect of contrastiveness on holistic, promoting relational perception at the spatiotemporal dimension. A context promoting a true balance for relational and figural perception, such as at the center of the scale in Figure 1.2, might promote holistic perception. It is also true that an implausible situation might keep the relational prominence to the highest, such that they will be

specifically recalled afterwards more than many normative contexts which are abundant in nature (Friedman, 1979). For example, while driving back home, someone might be more likely to report that, “I saw a dustbin on the crossing” than reporting normative situations like, “I saw a dustbin on the roadside” (unless it is really relevant). Hence, events evoking contrasts out of implausibility, thereby foregrounding experiential and relational aspects, are more likely to remain in attention than events which only bring figural attributes into attention (Hunt & Worthen, 2006). Although the example above depicts an ideal case from visually driven contrast, the phenomena can also be observed for other dimensions where relational aspects are highlighted. Learning from self-correction while solving a problem (let’s say while coding) is suitable for such analysis. Engaging with unsolved mistakes (contrasts arising from novel cases) is more likely to foster deeper learning through experiential engagements than mindlessly working on already solved solutions repeatedly (a situation akin to normative situations) (Marton & Booth, 1997; VanLehn, 1991). A deeper form of learning is also based on contrast, where multiple options are tested, verified, rejected, and eventually accepted, accumulating in experiential traces that are not achieved while working with solved problems alone.

Thus far, I have focused on *situative* contrasts, in which the surrounding context provides a background that supplies semantics for the entity in question. At a more fine-grained level, however, contrastive contextualization can also operate over *functions*—for instance, over the actions an environment affords (Gibson, 1979)—and over *locative* dimensions, where an entity is highlighted through its relational saliency against possible alternative realizations in the same context (Herskovits, 1986). Similar contrastive effects also arise for *events* that unfold along a continuous spatiotemporal dimension (Zacks & Tversky, 2001). At a broader level still, contrast structures social interaction: relational prominence is often achieved by invoking an alternative state of affairs, frequently marked linguistically through negation (Verhagen, 2005). Crucially, the aim of the thesis is to provide a universal continuum working for all such forms of contrasts, providing a foundation for comprehensive analysis based on relational prominence affecting contrastiveness, its perception and understanding regardless of the nature of contrast observed on any (contrastive) dimensions.

1.2 Dimensions

Consider the images in Figure 1.3. At first glance, they depict familiar objects (e.g., dustbin, hammer, ball, screws, agents) along with their visible attributes (shapes and color). However, beyond the objects and their attributes, we also see some relations emerging. These relations perhaps become more relevant than individual objects and their attributes in few contexts. These relations capture the between-objects properties rather than properties each object has on its own (e.g., within-object properties). Whereas within-object properties are symbolic, e.g., object attributes, between-object properties are

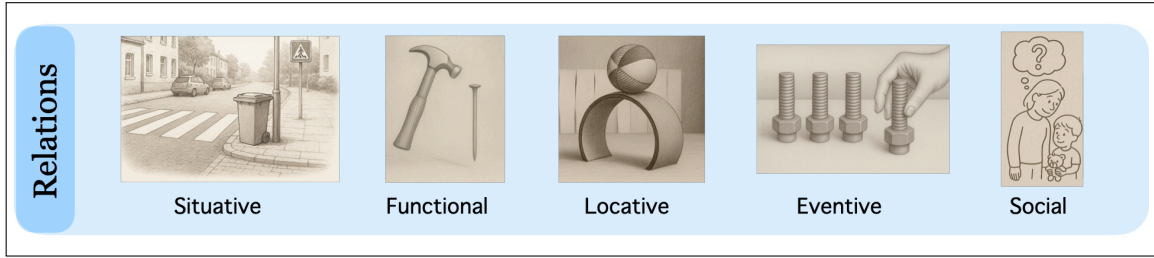


Figure 1.3: Dimensions of relations. A visualization of the five key domains where relational properties emerge: Situative (spatial configuration), Functional (learned affordances), Locative (exhibiting alternate locations), Eventive (spatio-temporal expectancies), and Social (social scripts).

experiential. Relational properties are experiential in nature, in that they are acquired through our interaction with the environment, and are driven by concrete representations rather than by abstract, attribute-like descriptions (Barsalou, 1999; Wu et al., 2014). This aligns with the view that experiential representations are action-isomorphic: they preserve the structure of what we do and perceive as we experience it, in contrast to symbolic representations, which are typically treated as more abstract (Kaup et al., 2024).

Our experiences establish a relational *ground* against which the entity in concern (i.e., figure) is perceived. For example, a situative relation is driven by normative experiences, such as (dustbin is in most cases found not anywhere other than roadsides) (Biederman, 1981). Similarly functional relations form a background for possible actions emerging from experiential traces or simulation of experiences (Gibson, 1979; Glenberg, 1997). The idea has already been dominant in the ecological framework, known as affordances, e.g., a nail affords to be hit by a hammer (Gibson, 1979; Norman, 1988). Once we identify the figure, its attributes, and the ground exhibited by experiential traces, contrastiveness can be manifested by the relational dynamics for all the provided relations, as shown in Figure 1.3. In this framework, I distinguish (i) *situative* dimension, which is driven by normative expectations about the situation and where the figure belongs in general (Biederman et al., 1982; Henderson & Hollingworth, 1999); (ii) *locative* dimension, where locative placement highlights an entity against plausible alternative placements (Herskovits, 1986).² (iii) *functional* dimension, which organizes possible actions in terms of affordances (Gibson, 1979). (iv) *eventive* dimension, where contrast emerges from how events unfold over continuous spatiotemporal structure (Zacks & Tversky, 2001); and (v) *social* dimension, where contrast is negotiated in interaction by orienting to an alternative state of affairs in joint action, often made explicit through linguistic resources such as negation, repair, and other contrastive constructions (H. H. Clark & Krych, 2004; Verhagen, 2005).

These dimensions need to be considered jointly because contextual interpretation can

²To distinguish locative from situative dimension, it is important to consider the level of analysis. Whereas situative contrast is driven by violations of experiential regularities, locative contrast is driven by spatial alternatives (Biederman et al., 1982; Henderson & Hollingworth, 1999; Herskovits, 1986), and can therefore be treated as a specific case.

rarely be expressed on a single relational axis, as we’ll see in chapter 3 (Contexts). In this sense, what becomes salient in a scene is typically the outcome of how situative, functional, locative, eventive and social constraints converge or compete in shaping Figure–Ground organization. Supplemented by relational analysis within the framework of contrastive contextualism, it can be demonstrated that the degree of contrast manifested along each dimension is sensitive to the distance between Figure and Ground—a notion I refer to as *relational prominence* (cf. Fig. 1.2). In the following section, I’ll first attempt to identify the figure and ground at each dimension supplemented by examples, and present few results from my past studies. Later, I will show how the contrastive and relational prominence can be used to identify the figure and ground dynamics on the contrastive continuum for most contextual interpretations in which figure is situated.

1.2.1 Situative Contrast

In Rohlfing et al. (2003), a *situation* is defined as concrete spatiotemporal ordering of objects and agents under physically given constraints (e.g., gravity or light), that is, the immediate configuration an agent can in principle perceive. Within this framework, situation is never interpreted in isolation, but is embedded in multiple, potentially overlapping *contexts* that constrain what counts as an appropriate analysis and what kinds of actions are licensed. Moreover, the contextual layer can be unpacked into an *intercontext*, the socially shared reality that emerges from patterns of interaction, and an *intracontext*, the agent-specific mapping abstracted from past situated experience. Thus an intracontextual structure provides the experiential *Ground* against which a current state of affairs is construed. Similarly, Gentner (1983), defines situation as systems of objects, *object-attributes* and *relations* between objects which function as wholes at a given level of organization.

In day-to-day scenarios, we are highly attuned to perceive figure and ground together, and hence rarely reason for figure’s existence against ground, in such cases the figure becomes our principle percept. This tendency is well documented in the event-cognition literature as a form of goal-biasness, where attention is preferentially drawn to the figure or principal entity (Lakusta & Landau, 2005; Regier & Zheng, 2007). More generally, figures in a situation induce systematic biases, such that the ground is less likely to enter the locus of attention because the figure is prominent (Gleitman, 1990; Mazza et al., 2009). Developmental findings further suggest that this intrinsic figure-bias emerges early, plausibly because goal prominence serves a functional role in learning to parse continuous actions and events (Hespos et al., 2009, 2010).

For the *situative* contrasts to occur, this default figure-driven perception must be disrupted where we are compelled to reason about the Figure *relative to* the Ground, typically because the figure’s presence violates what experience has made normative. In other words, situative contrasts emerge when we reason about the very existence of figure against a ground which are often resultant from our experiences. For example, imagine someone coming from a country which follows left-side driving rule and has never seen or known

right-side driving exists, when the person finds herself in a right-side driving, this would be perceived as contrast with ground, in such cases, roads and their rules becomes more prominent for that moment over individual vehicles. Conversely, given a situation not all contrasts might necessarily originate from our experiences, for example, assume we do not have any background or have known about earthen pots before, and travel somewhere to observe someone making them by hand for the first time. We will still reason relations and actions, such as the reasoning behind the existence of the pot in that particular situation (against ground) even though we do not have experience or seen pot making before. Meaning, when there is no experiential ground, the relational prominence can also be foregrounded. Conversely, for the potmaker, it is a regular thing and he or she might not reason about figure (pot) existing against ground, leading to no more relational prominence than figure itself.

In the dustbin example, one may now try to identify the components that are responsible for contrastive perception. Figure is the principal entity around which our default perception lies in a normative situation, such as *dustbin* in our current example. Defined by their attributes, figure also absorbs the meaning from its surrounding and highlight them differently how they find themselves against a ground. According to Talmy, “the Ground is a reference entity, one that has a stationary setting relative to a reference frame, with respect to which the Figure’s site, path, or orientation is characterized.” (Talmy, 1975, p. 184). The ground is hence formed of all the entities perceived holistically against the figure. There is a relation between figure and ground which is formed by experiential traces. How far the relational component lies between figure and ground establishes the degree of contrast. This can be visualized in the scale provided in Figure 1.2.

Taken this notion, almost all situative interpretation can be drawn in terms of contrast they reflect with respect to how figure and ground relate. Few examples are provided in figure 1.4. How far the figure and ground lie from experiential ground decides the contrastiveness. In this figure the arrow length depicts the magnitude of contrastiveness.

Placement on the contrastiveness scale. The contrastiveness scale introduced in Figure 1.2 provides a general continuum from highly normative to increasingly implausible Figure–Ground configurations. In the present subsection, I use this same continuum as a reference frame and illustrate it with situative cases in Figure 1.4, where each configuration can be read as occupying a different position on the scale. Situative contrast concerns how a Figure sits against its Ground, relative to the alternative placements that the layout visibly licenses. As the number and salience of credible alternatives increase, and as the realized placement approaches (or crosses) a categorical boundary, the relation itself becomes increasingly prominent.

Low situative contrast (small relational prominence due to routine placement). Most situations exhibiting no or low contrasts are defined by normativity, such that they are highly routinized (Mandler, 1984; Schank & Abelson, 1977). The situative layout

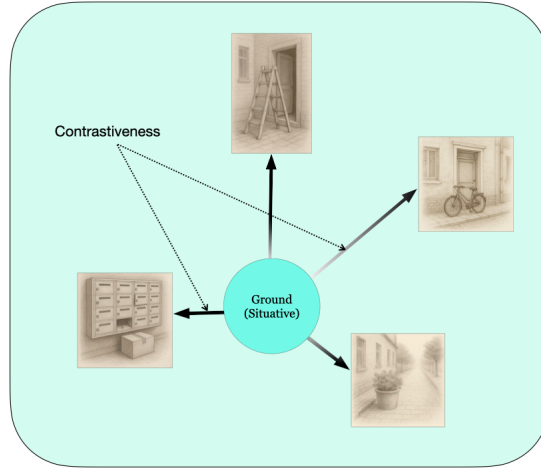


Figure 1.4: Examples of situative contrast. The central node represents the Ground (situative expectation), while the surrounding images depict Figures (realized situations) at varying distances from this ground. The length of the arrow corresponds to the magnitude of contrastiveness or relational prominence.

effectively offers a single, canonical location, so the relation feels “transparent” and unremarkable. In this case, the perceiver may register the object but does not necessarily dwell on the rationale for its existence in a situation, such as why the entity stands against the specific ground. This is because the placement does not violate experiences associated with the figure’s existence against the ground, and hence the relational prominence is low. An example of such situations could be: a trash bin parked by the sidewalk (a normal case) or a parcel lying near a mailbox grid, as shown in Figure 1.4.

Mild situative contrast (mild relational prominence due to mild implausibility). In this case, the perceived situation is acceptable yet slightly off the expected line. This “near-miss” character makes the relation mildly salient without reading as a violation: the viewer notices a small negotiation with the layout but still recognizes the intended placement as plausible (Rosch, 1975). In figure 1.4, such contrasts are recognized by situations with a distance lesser than high situative contrasts but greater than low situative contrasts, standing somewhere in the middle where the relations are not automatically foregrounded but become apparent when consciously reasoned upon. Examples of such situations are shown in figure 1.4 with medium-length arrows from the ground, such as a planter nudged outward from the pedestrians but still leaving the pathway clear. Similarly, a bike parked diagonally near a doorway leaving a clear passage. These configurations surface a possible implausibility based on experiences (“pathways should be generally clear”) but do not force an explicit implausibility (“planter can be on the side of pedestrians”).

High situative contrast (large relational prominence due to high implausibility). These situations exhibit the highest contrast, where the relations are maximally foregrounded even before the isolated attributes. In these cases, the relations dominate the perception, where the scene is more likely to be described by the relations it exhibits than by individual entities. Highly contrastive situations stand on implausibility driven by our experiences, where the figure’s existence against a ground is reasoned (Biederman et al., 1982; Friedman, 1979). Examples of such highly implausible situations leading to contrast are: a trash bin placed in the crosswalk, foregrounding “on the path”; a bench set across the crosswalk, visibly obstructing the traverse; a bike parked across a doorway, competing with entry/exit access. In these scenes, the layout advertises a canonical alternative, and the realized placement pointedly opposes it, producing strong relational prominence.

1.2.2 Functional Contrast

Though the majority of relations are spatially construed within a larger ground, functional relations are in themselves a more specific case due to their dependence on possible actions afforded by the entities concerned. Grounded in sensorimotor theories, the perception of such relations has already been highlighted in ecological psychology (Chemero, 2009; Gibson, 1979). An example of such a relation is provided in Figure 1.3.

Whereas situative contrasts foreground an entity against a broadly construed scene, functional contrasts seem to foreground the relation between two (or more) entities by virtue of the actions they afford together. In this sense, contrast is ubiquitous for functional relations; that is, every tool, container, or support enters a context already indexed to a family of practicable routines (hammer–nail, pitcher–glass, key–lock). What makes the contrast perceptible is not the mere co-presence of objects, but the relational dynamics of afforded actions grounded in our experiential traces. When the enacted or implied action departs from what prior experience prepares us to expect, the functional ground (i.e., the action schema binding the entities) becomes more prominent than the isolated perception of the figure’s attributes, raising contrastiveness on the continuum (cf. Fig. 1.2).

Function is thus a relational primitive anchored in experiences rather than isolated perception attached to object attributes. In line with this, previous studies have shown that functionally interacting pairs are perceptually grouped and identified more efficiently than the same objects arranged without a workable interaction; disrupting the interaction impairs identification, revealing that cognition privileges the action relation over the stand-alone object when the two come apart (Green & Hummel, 2006; Riddoch et al., 2003). This pattern aligns with the present view that increasing functional mismatch increases relational prominence and, hence, perceived contrast.

In functional contrast, the figure is the action-relevant entity in focus (e.g., the hammer), while the ground comprises the functional relations grounded in experiences that make a particular action possible. Crucially, these relations are manner-driven: striking

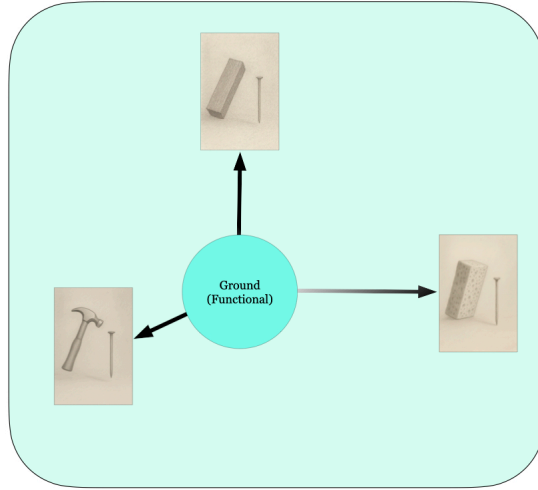


Figure 1.5: Examples of functional contrast. The central node represents the Ground (functional expectation or affordance schema), while the surrounding images depict Figures (realized actions) at varying distances from this ground. The arrow length depicts the degree of contrast from the canonical function.

vs. tapping, pouring vs. dribbling, cutting vs. scoring (Talmy, 1985a). Manner refines the actions and therefore the degree to which the realized relation fits backgrounded experience. As manner departs from the expected routine (e.g., a saw used with a dragging rather than cutting motion), functional prominence rises and attributes recede.

Placement on the contrastiveness scale. Functional contrast asks how well the interaction between figure and ground fits a practiced affordance schema (what the ground “invites” the figure to do) (Gibson, 1979; Norman, 1988). As with situative cases, the scale is the relational distance between the observed use and the canonical use (Tversky, 1977). This means the greater the distance, the stronger the relational prominence (here we notice the relation more than the individual objects).

Low functional contrast (small distance with routine use). The realized action cleanly instantiates a familiar schema with the expected roles and forces. Execution is efficient, controllable, and aligns with prior experience, so the relation feels “transparent” and attention stays on the task outcome rather than on how the objects are being used (R. Cooper & Shallice, 2000). A few examples of such functional relations are: a hammer driving a nail into wood, a key turning a lock, a mug used for pouring water into a glass, or a knife slicing bread on a cutting board. Essentially, these relations are supported by perceptual cues which justify the ground’s affordance without strain (Riddoch et al., 2003).

Mild functional contrast (moderate distance with workable but non-canonical). Here, the action accomplishes the goal with action format substitution or role stretching.

Meaning, the affordance is partially “canonical”, grip, or contact is adequate but not ideal, and hence the relation becomes mildly salient (Osiurak et al., 2010). Observers register that the action works, but “it isn’t what this thing ideally is for.” Examples of such relations include using a sturdy wooden block or shoe to tap a nail, opening a paint can with a butter knife instead of a flathead screwdriver, using a phone as a paperweight, turning a screw with a coin, or propping a door with a backpack. Here, the ground’s affordance is partly satisfied, but alternative canonical tools remain salient competitors (Buxbaum, 2001).

High functional contrast (large distance with schema breach). For high functional contrast, the observed interaction conflicts with the learned affordance; i.e., contact mechanics, material properties, or role assignments make the action ineffective or self-defeating (Bach et al., 2009). Here, the relation itself dominates perception before the individual features can offer an interpretation, and the scene reads as wrong or non-canonical. Examples of action mismatches under this format include trying to drive a nail with a sponge, a tomato, or a sheet of paper; attempting to cut bread with a paintbrush; or stirring soup with a fork made of chocolate. In this case, observers infer that the affordance schema is violated, not merely stretched as in mild functional contrast (Sitnikova et al., 2003).

Functional relations are ever-present and affordance-grounded, and contrast emerges when the afforded action diverges from experiential traces (Barsalou, 1999). The more the realized interaction (including its manner) deviates from the backgrounded schema, the farther the relation moves from the figure on the contrastiveness continuum, and the more cognition privileges functional ground over entity attributes in making sense of the actions. This logic situates functional contrast within the same unifying scale used for situative contrast, extending the notion that contrast is a contextual phenomenon rooted in relational prominence.

1.2.3 Locative Contrast

Just as events have a temporal background, locative contrast has a spatial background (Jackendoff, 1983). This means that a scene offers a set of credible alternate placements for a figure with respect to a ground, and the present placement is interpreted against those alternates. Thus, when a figure is “on” something, it is simultaneously not under it (Figure 1.3, Locative). For example, as shown in Figure 1.6, if a ball is placed on a ring, most people will also interpret the location as “not” under the ring. Meaning, when a ball is outside a box, the option of being inside is live. These oppositions (up–down, in–out, on–under, front–behind, left–right) are not “add-ons” but part of how the scene is holistically organized for perception (Landau & Jackendoff, 1993; Talmy, 2000). When such alternates are available, the relational component becomes prominent and the figure’s mere attributes recede. Conversely, when no alternate is licensed by the layout, relational

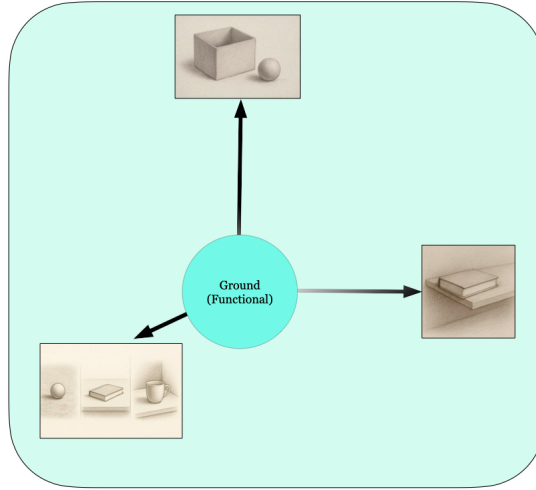


Figure 1.6: Examples of locative contrast. The central node represents the Ground (locative expectation or available set of placements), while the surrounding images depict Figures (realized placements) at varying distances from this ground. The arrow length depicts the degree of spatial competition or deviation from the canonical placement

prominence is weak; e.g., a ball on bare ground with no support or container in view does not invite a contrast because there is nowhere else within the scene for the ball to plausibly be. Under this view, the ground is the reference region that makes multiple placements possible, and the figure is the movable entity whose location is at issue (Herskovits, 1986). Locative contrast is therefore availability-sensitive: the more clearly the scene offers structured alternatives, the more the figure-ground relation comes forward. For instance, geometric cues, height differences, visible openings, edges, and support surfaces act like “spatial affordance markers” to project where the figure could go next and against which its current placement is read (Carlson-Radvansky, 1999). As with other contrasts, this effect seem to be graded where subtle misalignments may only bring the relation into view, while placements that sit on the boundary between roles (e.g., balanced on a rim) can make the relation a principal percept (Regier, 1996).

Placement on the contrastiveness scale. Locative contrast is about how a figure’s placement stands out against the set of credible alternatives the scene makes available. In all three levels below, the “distance” in the scale is the relational distance between figure and ground, i.e., how many live alternatives the layout licenses, and how strongly the realized placement competes with the canonical one (Herskovits, 1986).

Low locative contrast (small distance with no alternates). In this situation, only one plausible placement is effectively licensed by the layout, so that the relation reads as transparent and unremarkable. Here, attention falls on the object’s attributes because the spatial relation does not provide alternates or perform any explanatory work

(Talmy, 2000). In figure 1.3, such relations can be seen for a ball resting on bare ground with no container or support in view, a book lying flat on an otherwise empty tabletop, or a cup seated on a single visible counter where no other support is present.

Mild locative contrast (moderate distance with gentle spatial competition). When the scene offers more than one reasonable placement, and the realized placement is acceptable but the alternate does not entirely compete for the figure’s location, a mildly locative contrast emerges. Crucially, here the alternates are visible enough to enter awareness such that they compete for the figure, making the relation mildly salient. We can think of such contrasts as almost “canonical” placements (Carlson-Radvansky, 1999). Examples of such contrastive relations are also seen in figure 1.3, such as a book resting near the shelf’s edge (“plainly on,” but the location “to the floor” is a live alternate), or similarly, a cup on the table with the floor below, i.e., where other supports are present but weak.

High locative contrast (large distance with strong spatial competition). For high locative contrasts to emerge in a scene, an alternate or multiple salient placements should be available. Here, the realized position either competes head-to-head with a canonical alternative or gets its definition through an alternate possibility, just as down is always relative to up or on is always relative to under (Singh & Rohlfing, 2025a). When the relation itself becomes the main percept, we notice “where” before “what” (Landau & Jackendoff, 1993). Such contrasts are ideally characterized by negations reflecting alternates, such as saying “not up” implies “down” in a situation where both alternates are presented as locative contrasts. In our past studies, we utilized locative contrast to show that such contrastive situations can be verbally guided through negations (Singh & Rohlfing, 2025a). An example of such a contrast is a ball perched on the rim rather than inside or under it. enditemize

Following the above descriptions, two boundary conditions for the contrastive scale should follow. First, the absence of alternates exhibits minimal contrast, so that without any credible locus, the relation remains backgrounded and attributes associated with the figure dominate. Second, more structured alternatives provide stronger contrast through alternate possibilities of where the figure could have been, for instance, adding a shelf above a table or introducing a visible edge increases relational salience by making the other placements live options (Gibson, 1979). In all such cases, what we perceive first is not the object in isolation but how it is spatially bound to its ground relative to the alternates the scene makes available, and the configuration moves along the contrastiveness scale accordingly.

1.2.4 Eventive Contrast

Events unfold on a spatiotemporal dimension. In this view, eventive contrast arises when the present sequence of an event is interpreted against a history that the scene

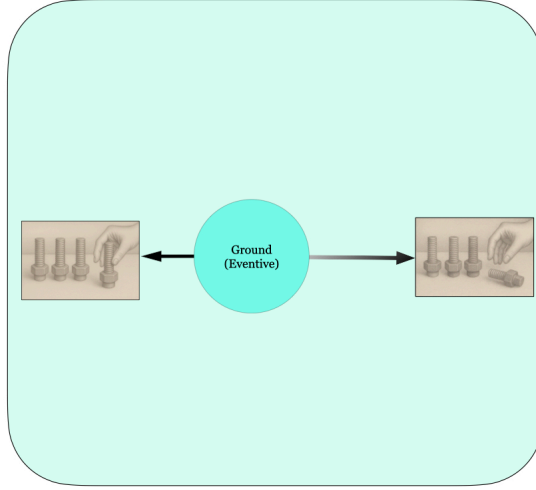


Figure 1.7: Examples of eventive contrast. The central node represents the Ground (eventive expectation established by history, e.g., a repeating pattern), while the surrounding images depict Figures (realized next actions) at varying distances from this ground. The arrow length signifies the degree of deviation from the projected sequence

itself has already made available (Zacks & Tversky, 2001). Here, the past is not backgrounded as mere memory, rather, it is part of the perceivable arrangement with regularities, rhythms, and partial trajectories projecting the next move. When a sequence establishes a pattern, it is tempting to think that the next occurrence should read through that pattern. In cases where the realized action diverges, the *relations* emerging holistically become prominent over the attributes of the single subevent in hand. In other words, the event history furnishes the ground against which the current subevent, the *figure*, is interpreted. In this sense, the contrast should emerge as a mismatch between the expected continuation and the actual continuation of the ground events. Without an available history (no evolving scene), or with a regular continuation (repeated event), eventive contrast has little to grip and will not be strongly perceived. Concretely, consider the screws example from Figure 1.7: when the prior screws are placed vertically in a tidy row, the upcoming screw (the figure) is prepared, and the background rhythm licenses the expectation “vertical again” (Singh et al., 2025). A horizontal placement now constitutes an eventive deviation such that the manner’s relation (past-present) is foregrounded over the current move. Similar spatiotemporal patterns are also observed in other situations, e.g., three boxes slid leftward one after another where the fourth is pushed forward instead. In each case, the ground is the unfolding regularity, such as the partial pattern that projects a near-future, and the figure is the present action that either continues or disrupts it. Eventive contrast is therefore also manner-sensitive; i.e., not only what action follows but how it is realized can align with or contrast with the projected course.

Placement on the contrastiveness scale. Eventive contrast is the temporal counterpart to the spatial and situational cases discussed earlier. It arises when the present subevent is read against a history that the scene itself has already made available (Zacks & Tversky, 2001). In this view, the unfolding pattern, i.e., partial trajectories, spacing, rhythm, orientation—constitutes the *ground*, and the current subevent is the *figure*. What we call “distance” on the contrastiveness scale is thus how far the current action departs from the projection licensed by that ground (Reynolds et al., 2009). When the projection is weak or absent, departures cannot be registered and eventive contrast remains low. Conversely, when the projection is strong and regular, even small deviations surface as salient relations. This framing ties eventive contrast directly to relational prominence as defined for the other dimensions.

Low eventive contrast (no deviation from ground). At the low end of eventive contrast, the present action simply continues the established sequence in both *what* it does and *how* it is realized and the figural-event ratifies the ground. For example, a fourth screw is placed vertically with the same spacing and angle as the first three, or a painter’s fourth stroke repeats the same top-to-bottom motion at the same tempo (Singh et al., 2025). Such continuity keeps the relation transparent and the pattern holds, where the projection and the realization coincide, and attention is free to settle on outcomes (goal) rather than on the relation between move (manner) and history.

Mild eventive contrast (slight deviation from ground). A step higher, mild eventive contrast appears when the action preserves the overarching schema but shifts the manner variables. Here, the ground remains a reliable predictor at the level of path or role, but the figure negotiates the details, for example, the screw is still vertical yet slightly misaligned or spaced differently, the fourth step follows the same path but at a slower cadence, or a slider moves leftward as expected but with a brief hesitation or smaller amplitude (Singh et al., 2025). These contrast scales are “near-miss” continuations. They render the relation gently salient without reading as violations such that the projection stands, but the system now encodes how the realization departs from the routine. Here, since the manner is the locus of difference, contrastiveness rises only moderately.

High eventive contrast (maximum deviation from ground). At the high end, eventive contrast is produced when the present move cancels the projected endpoint altogether or substitutes an incompatible action, thereby collapsing the live script (Schank & Abelson, 1977). After lifting the kettle to pour, the agent returns it untouched, or after three hammer strikes toward driving a nail, the fourth move is to wipe the nail with a cloth. Here, the figure does not merely deviate, rather, it withdraws commitment to the projected goal or swaps scripts (Baldwin et al., 2001). The ground remains vividly present (we *know* what should have come next), and it is precisely this presence that renders the deviation maximally contrastive.

For eventive contrasts, we can imagine two conditions that should be followed. First, no history with minimal eventive contrast, such as an isolated act with no visible history offers little relational prominence because there is no credible projection for the figure to follow or violate. Second, the more stable the projection, the stronger the contrast can be. Factors such as repetition, clear partial trajectories, temporal rhythm, and evenly spaced placements strengthen the background expectation and thereby amplify the salience of departures (Kurby & Zacks, 2008). In sum, eventive contrast scales with the distance between the realized move and the pattern it stands against, with low contrast for ratified continuations, mild contrast for manner shifts within the same schema, and high contrast for cancellations or script substitutions. As other dimensions, the graded pattern preserves the idea of the contrastiveness scale (figure 1.2).

1.2.5 Social Contrast

Social contrast arises when an ongoing state of affairs involving agents is read against a normative background of social scripts (Schank & Abelson, 1977). In everyday interaction, the ground is not only spatial or temporal but normative; i.e., shared routines about what typically happens next (queueing, greeting, turn-taking, paying sequences). Against this ground, the present configuration of agents and actions (the figure) is interpreted. When the current action coheres with the script, the relation is transparent, and attributes (who, where, and what) dominate. Conversely, when the current action deviates from the normative script, the relational component becomes prominent, and contrastiveness rises (Garfinkel, 1967).

In this sense, social contrast is ubiquitous, where every social scene is already indexed to a set of practicable routines, and we only perceive the contrast when the realized action misaligns with those routines. Such social relations were previously defined in terms of concepts like help, hinder, chase, give/take, lead, or follow (Hafri & Firestone, 2021; Heider & Simmel, 1944). These states of affairs are extracted rapidly and with structured roles (e.g., Agent/Patient), so that a small deviation in who does what to whom can reorganize the percept from “normative” to “contrastive” even before deliberation sets in. This crucially matches the broader insight that sophisticated relations, including social interactions, are not just inferred but are seen with role structure and contributions. Consequently, departures from the norm make relational prominence salient over the individuals’ attributes (Hafri & Firestone, 2021).

At this point, it is useful to specify what counts as figure and ground in social contexts by treating particular social moves as figures against the scripted routines that sustain expectations as ground. For example, consider a social context such as a restaurant with a social move (e.g., the server presenting a bill now). Here, the ground can be considered as the scripted sequence and local cues that license expectations (e.g., menus → order → eat → bill) against whose schema the figure is interpreted (bill

presentation). The manner refines norm conformance; i.e., not just what action occurs but how it is done (timing, orientation) modulates prominence. A polite interruption at the right transition point may be weakly contrastive, whereas the same content delivered abruptly mid-course is more contrastive because the manner clashes with the script’s rhythm (Goffman, 1967). As in other domains, when manner and role-assignment conflict with the expected template, the relation itself becomes foregrounded.

Placement on the contrastiveness scale. Social contrast follows the same figure–ground logic as the other dimensions, with *normative scripts* providing the ground (shared routines about “what happens next”) and the *present social move* functioning as the figure (Schank & Abelson, 1977). Distance on the contrastiveness scale reflects how far the current act departs from the script that the scene itself licenses (queueing conventions, service rituals, turn-taking, authority handoffs). When the script is strong and unambiguous, even small departures become visible relations, whereas when the script is weak or absent, departures fail to register as contrast (Garfinkel, 1967).

Low social contrast (small distance). The present act continues the script both in content and in manner such that the relation remains transparent and attributes dominate. An example is the typical restaurant exchange where the actions proceed in order—*menus* → *order* → *eat* → *bill*—so the handover of the bill after dessert feels almost invisible as a relation (Schank & Abelson, 1977). Similarly, when a newcomer joins the back of a clearly marked line, acknowledges the host when addressed, and waits for a turn, the social ground predicts each move, and the figure merely ratifies it. In such cases, observers register who is present and what they are doing but do not dwell on the relation to the script because nothing in the act competes with it.

Mild social contrast (moderate distance). An act in a social context that shifts the timing or manner yet remains compatible with the underlying script, therefore still guiding expectations, can be considered a mild social contrast. For instance, in the restaurant context, when a server places the bill on the table slightly early, say, soon after the mains arrive. Here the billing step is brought forward, but its manner keeps it within bounds. Another example showing mild social contrast can be seen in a queue: a customer stands a little ahead of the floor marker without blocking anyone; here, the role and order are preserved, but the local alignment is off. Although these are near-miss continuations, the alternatives are live but weak, and the relation comes forward gently because timing and tone negotiate with the script instead of replacing it (Goffman, 1967).

High social contrast (large distance). The act reorders steps or swaps roles in ways that compete with the script, making the relation itself the principal percept. A striking case is a server presenting the bill before menus arrive such that the billing relation is pulled far forward, and consequently, the sequence, rather than the server’s

or diner’s attributes, becomes what is seen. Similarly, if a passerby cuts into a queue while meeting others’ gaze as if entitled, the visible role assignment (who yields to whom) is inverted. In institutional settings, a visitor issuing demands without greeting in a formal meeting, or a patron in a quiet library calling out across the room, all these moves collides with a salient norm and reorganizes authority and turn-taking (Milgram, 1992), and more often repairs (Schegloff et al., 1977). In such scenes, the ground advertises a canonical alternative so clearly that the realized move reads as a deliberate opposition to it, yielding maximal relational prominence.

As with other dimensions, no credible norm means minimal contrast. In unfamiliar or underspecified settings, an odd act may not read as contrastive because the ground is weak. Conversely, the stronger the scripted background, the stronger the contrast, such as clear turn-taking cues amplify the salience of departures. Because social relations are represented with abstract roles and are rapidly extracted, even subtle mismatches (e.g., who initiates, who yields, timing of handoff) can raise contrast by shifting role structure rather than object properties, aligning with the broader claim that we often perceive how agents relate before we attend to who they are (Hafri & Firestone, 2021). In sum, social contrast is the normative counterpart of situative, locative, and eventive contrasts, as it is grounded in shared scripts and role structure and become perceptible when the present move misaligns with that normative ground. It also scale with the distance between the realized act (figure) and the scripted sequence (ground) on the contrastiveness continuum as visualized in figure 1.2.

1.3 The multimodality of contrast

Having laid down the dimensions of contrast (situative, functional, locative, eventive, social), one important gap remains on how language may modulate the perception of these contrasts. Contrast being defined as nonlinguistic from the beginning, a motivation of such effects can be derived from research on language and thought (Gentner, 2003; Gleitman & Papafragou, 2005; Wolff & Holmes, 2011). Up to this point, the analysis has been largely modality-first, where we have treated scenes, actions, and sequences as contrastive because experiential traces provide a ground against which a figure is interpreted. Yet everyday cognition is not unimodal. Perception is routinely coupled with linguistic framing, verbal labels, discourse cues, and interactional routines. The question, then, is not whether contrast exists without language (it does), but how language can reshape, amplify, or attenuate contrastive perception once the relational aspects are in place. The current chapter is motivated by the idea whether and how language may act as a contrastive mechanism with other forms of contrast.

Multiple hypotheses in language–thought research characterize how language may influence nonlinguistic conceptualization (Wolff & Holmes, 2011). From this perspective,

three possibilities are at stake. First, a *language-first* account treats contrast as primarily a linguistic construction that can subsequently guide perception. Second, an *experience-first* account treats contrast as arising from action and history, with language largely reflecting and packaging what is already available from cognition. Third, a *hybrid* account holds that language tunes the use of contrasts that are already grounded in general cognition and experience.

According to the contemporary work on linguistic relativity (e.g., Wolff & Holmes, 2011), most evidence supports the hybrid view, i.e., language does not overwrite core perceptual/conceptual distinctions, but language-specific processes in preparation can reflect upon attentional patterns (“thinking for speaking”), meddle with ongoing non-linguistic processing, spotlight entities, augment representational resources, and induce schematic modes of analysis that persist into nonlinguistic tasks. These mechanisms may provide a systematic investigation route by which language may alter relational prominence and thus contrastiveness, i.e., the perceived distance between figure and ground on the contrastiveness scale.

I lay down here the above-mentioned possibilities and then discuss how these possibilities might reflect upon contrasts. The first possibility is “thinking-for-speaking”. When preparing a description, speakers align attention with what their language routinely encodes (Slobin, 1996). For example, in past studies it was shown that speaker of satellite-framed languages (where PATH information is encoded in additional constituent), while preparing to describe the scene are more likely to attend to the path and manner than speakers of verb-framed language (where MANNER information is encoded in the verb) (Papafragou et al., 2008). However, the effect receded when participants simply watched the motion events, without needing to describe the scene. Meaning, language affects event conceptualization only when linguistic representations are activated as a result of verbal encoding. Such a processing pathway may affect contrastive perception. For example, language can determine which relational alternate becomes salient only when a relation is verbalized, thereby moving a configuration rightward or leftward on the scale (figure 1.2) without changing the scene itself.

The second possibility is that language may act as a “meddler” for contrastive perception (co-active influence). Verbal labels and constructions can make particular distinctions hard to ignore and thereby bring relevant units under attention (Lupyan, 2012). For example, in Singh and Rohlfing, 2025a, it was shown that using verbal labels before or after concurrent action modulate gaze shift toward the goal of the action, potentially suggesting a co-influence of language on action conceptualization. Crucially, when the instructions contain negations to negate the current event in light of the previous event in contrastive actions, the event memory was better, potentially highlighting the higher relational prominence. Here the relation drew attention because the verbal cue (negation) brought the current and the already happened event within a unified frame of interpretation.

The third possibility is that language acts as a spotlight. Under certain conditions it is valid to assume that language can act to highlight specific properties and may act as a spotlight, making certain aspects of the world more salient than other (Vigliocco et al., 2005). Few language specific regularities and rules developed after habitual use of language might make particular aspect of the world more salient than others. For example, it was found that grammatical gender learned within a language (such as through proper nouns) affects the association formation of that particular proper noun with common noun when the target common noun does not conform to the gender of proper noun (Sera et al., 2002). Regarding the relational perception, in a cross cultural study with Korean and English speaking children, it has already been shown that by 3 years English speaking but not Korean speaking show reduced sensitivity to tight and loose fit distinction, a relational feature prominent in Korean but not in English (Choi et al., 1999; Loewenstein & Gentner, 2005). This suggests that learned regularities in the language might spotlight specific constituents useful for nonlinguistic encoding. Hence, if a language in itself exhibit some contrastiveness through construction it may have influence how the relational prominence which may be spotlighted in a given context.

The fourth possibility is that language act as an augementer. As an augementer the language may provide new resources complementing the non-linguistic processes (Frank et al., 2008). Under this view, linguistic structure can extend nonlinguistic representations, such as it can enable exact rules, embed propositions, or create stable relational categories through some linguistic means (Lupyan, 2005). For example, it was found that during a nonlinguistic task requiring linguistic strategies to solve it efficiently, participants preferred to represent the problem as linguistic task even though the task provided an opportunity to be solved through visual simulation such as in “gear turn task” (Schwartz & Black, 1996). Thus given a situation language can augment contrastiveness if it may appear befitting for perception. In other words, in a contrastive frame, language can create additional plausible alternates (new candidate grounds or manners) that were previously unavailable to the system, thereby increasing the space of contrasts a scene can support.

And the fifth possibility is that language might act as an inducer (prime a particular mode). Meaning, language can set a stage for particular mode of processing that continues to be engaged even after language is no longer is in use (Boroditsky, 2001). This processing path has an important consequence for contrastive perception since it was found that relational language promoted mental simulation even when people describe a completely different scene. In this sense, using relational words can push perception toward schematic, relation-first processing that carries over to subsequent nonverbal tasks. In these terms, inducer effects raise the baseline salience of relations, making it more likely that observers will report the grounded configuration (e.g., ON, IN, ACROSS) rather than the object’s attributes, even when no explicit description is

required.

In conclusion, a question needed to be addressed in future is that how the linguistic influence of nonlinguistic processes can be defined on the contrastiveness scale as distance between figure and ground? From the above synthesis it can be postulated that, language can (a) alter which ground is selected (e.g., by spotlighting one relational aspect over another), (b) change the granularity of available alternates (augmenter effects, adding categories and roles), and (c) modulate the baseline salience of relations (inducer and spotlight), thereby shifting a display along the scale without changing the underlying scene. In practical terms, verbal cues can move a contrastive perception from low to mild, from mild to high (explicitly opposing the canonical template). If contrast is a general relational phenomenon, and if language can shift relational prominence through “thinking for speaking”, spotlighting, meddling, augmenting, or inducing, then a systematic account of language as a contrastive mechanism along the enlisted dimension is warranted. The next chapter develops this account, by providing examples of how linguistic devices (e.g., negation, contrastive connectives, prepositions, aspect) most effectively reconfigure the ground, stabilize alternates, or amplify mismatch, and how these effects register on the contrastiveness scale across situative, functional, locative, eventive, and social dimensions.

1.4 Roadmap and contributions

The central aim of this dissertation is to integrate empirical, methodological, and modelling contributions to the study of contrastive guidance in interactional contexts. The structure of the dissertation follows a conceptual progression: it begins with contrast as a Figure–Ground relation (Chapter 1), develops a formal model of contrastive guidance (Chapter 2), extends this account to interactional context (Chapter 3), and finally examines scaffolding and proposes an interactive extension of the model (Chapter 4). The publications on which the dissertation builds are mapped to the respective chapters below. A more detailed mapping between the dissertation chapters and the underlying publications is provided in Appendix A.

1.4.1 Publications and summary

The dissertation draws on the following publications and conference contributions. Together, these works develop the concept of contrastive guidance across a sequence of experimental studies, extend it to interactive human–robot settings, and examine its role in situated language use. In particular, they investigate how contrastive verbal guidance, especially negation, shapes visual attention, event memory, and repair during real-time language processing in dynamic contexts. The contributing publications are summarized chronologically below.

Contrastive Verbal Guidance: A Beneficial Context for Attention To Events and Their Memory?

Singh, A. and Rohlfing, K.J. (2025), Contrastive Verbal Guidance: A Beneficial Context for Attention To Events and Their Memory? *Cognitive Science*, 49: e70096.

Summary This article examines whether negation can function as a form of contrastive verbal guidance in action demonstration tasks. Across two eye-tracking experiments, participants observed contrastive and noncontrastive motion events while receiving assertive or negative verbal instructions either before the action or synchronously with it. The findings show that verbal guidance modulates attention to the action goal and that assertive–negative instructions particularly support the recall of contrastive actions. The study further reveals a trade-off between goal-directed attention and event memory. Participants recalled actions better when their attention was less strongly biased toward the goal. The paper therefore provides the central empirical foundation for this dissertation by demonstrating that contrastive guidance can de-bias attention from the most salient event component and support a more holistic representation of dynamic events.

Contrastiveness in the Context of Action Demonstration

Singh, A., & Rohlfing, K.J. (2023). Contrastiveness in the context of action demonstration: an eye-tracking study on its effects on action perception and action recall. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45, Sydney, Australia.

Summary This paper investigates how visual and verbal contrast jointly shape action perception and event recall during action demonstration. In an eye-tracking experiment, participants observed motion events that were either visually contrastive, such as Up–Down or Down–Up, or visually non-contrastive, such as Up–Up or Down–Down, where the accompanying verbal guidance varied between assertive and negative formulations. The findings show that visual contrast alone did not improve memory; rather, recall was facilitated when a visually contrastive action was paired with contrastive verbal guidance, especially in the assertion–negation condition. The paper contributes to this dissertation by showing that negation can reduce processing costs and support event memory when it is grounded in a prior positive context, thereby establishing contrastive guidance as a multimodal mechanism for structuring event understanding.

Coupling of Task and Partner Model

Singh, A., & Rohlfing, K.J. (2024). Coupling of Task and Partner Model: Investigating the Intra-Individual Variability in Gaze during Human–Robot Explanatory Dialogue. *26th ACM International Conference on Multimodal Interaction (ACM ICMI 2024)*, Costa Rica.

Summary This paper examines the coupling of task and partner models in human–robot explanatory dialogue. The task model was operationalized through two action components, path as the “where” dimension and manner as the “how” dimension, while the partner model was operationalized through gaze-based indicators of whether the robot was treated as a capable scaffolding partner. In an interaction with a NAO robot, participants placed objects according to the robot’s instructions and asked clarification questions when either path or manner information was missing. The results show that path-related questions were asked earlier than manner-related questions, supporting the idea that path is more prominent in action conceptualization. Moreover, participants primarily looked at the robot’s face during information-seeking, but shifted more toward the torso when the robot failed to provide a resolution, suggesting an exploration of the robot’s capabilities. This contribution extends the dissertation from controlled action perception to situated explanation by showing that task-related attention and information-seeking are partly shaped by the evolving partner model in interaction.

Manners Matter: Action History Guides Attention and Repair Choices During Interaction

Singh, A., Wrede, B., Richter, B., Groß, A., & Rohlfing, K.J. (2025). Manners Matter: Action history guides attention and repair choices during interaction. *IEEE International Conference on Development and Learning (ICDL 2025)*, Prague. (Talk)

Summary This paper investigates how action history shapes attention and repair behaviour during human–robot interaction. In an HRI experiment, participants followed instructions involving different action manners, such as placing objects vertically or horizontally, while the robot occasionally omitted manner information. The study manipulated the preceding action history to create either simple situational models, in which only one manner had been established, or rich situational models, in which competing manners were available. The findings show that rich situational models led to increased attentional engagement and more restricted repair requests that targeted the specific missing manner information, whereas simple situational models were associated with more open repair strategies. This contribution extends the dissertation by showing that contrastive guidance is not limited to isolated verbal instructions, but emerges

from action histories and contexts that structure attention, constrain possible interpretations, and shape how participants linguistically repair uncertainty in interaction.

Embedding Psycholinguistics: An Interactive Framework for Studying Language in Action

Singh, A., & Rohlfing, K.J. (2025). Embedding Psycholinguistics: An Interactive Framework for Studying Language in Action. *6th Biannual Conference of the German Society for Cognitive Science (KogWis 2025)*, Ruhr-University Bochum, Germany. (Talk)

Summary This contribution presents a conceptual and methodological framework for studying psycholinguistic processes online as they unfold in situated interaction. It argues that language, vision, and action should not be treated as isolated modules, but as dynamically coupled processes that mutually shape one another in real time. Using a human–robot interaction setup with eye-tracking, the study examines conversational repair as a window into this bidirectional language–vision–action interplay. Participants followed object-placement instructions from a humanoid robot and initiated repairs when relevant action information was omitted. By manipulating simple and rich situational models, the study shows that repair formulation and visual attention are jointly shaped by the available action context: restricted repairs were associated with longer dwell times, particularly in rich contexts where competing action possibilities were visually available. This contribution positions the dissertation within an interactive psycholinguistic framework, showing how controlled experimental paradigms can be embedded in naturalistic interaction to study the real-time coupling of language, attention, and action.

Scaffolding the Human Partner by Contrastive Guidance in an Explanatory Human–Robot Dialogue

Groß, A., Singh, A., Banh, N. C., Richter, B., Scharlau, I., Rohlfing, K. J., & Wrede, B. (2023). Scaffolding the human partner by contrastive guidance in an explanatory human-robot dialogue. *Frontiers in Robotics and AI*, 10.

Summary This article investigates whether negation can be used as a contrastive guidance strategy in explanatory human–robot dialogue. In a touchscreen-based task, a virtual robot guided participants to perform object-related gestures, while instructions were presented either contrastively, by affirming the intended manner and negating an alternative, or non-contrastively, by only affirming the intended action. The findings show that contrastive instructions initially increased processing time, reflecting

the known processing cost of negation, but also improved task performance: participants required fewer attempts and produced gestures that more closely resembled the demonstrated target gesture. Participants also rated contrastive explanations as more helpful than non-contrastive ones. This publication contributes to the dissertation by showing that contrastive guidance can be used to scaffold manners in human–robot interaction, where negation not only marks what should be avoided but also helps the human partner select and execute the intended manner more precisely.

A methodological framework for studying online language–vision interplay in human–robot interaction

Singh, A., & Rohlfing, K.J. (2025). A method for studying online language–vision interplay in a human–robot interaction. *Journal of Psycholinguistic Research*. under review.

Summary This manuscript develops a methodological framework for studying online language–vision interplay in human–robot interaction. It argues that traditional visual world paradigms offer strong temporal precision for studying language-mediated eye movements, but often remain limited by screen-based, unidirectional, and non-interactive designs. To address this limitation, the manuscript proposes human–robot interaction as a controlled yet interactive test bed in which comprehension, production, action, and gaze can be studied within the same unfolding task. Using a proof-of-concept implementation with a NAO robot, physical objects, eye-tracking, and repair sequences, the paper shows how gaze can be analyzed in relation to referential integration, partner-directed monitoring, and pre-utterance planning. The empirical illustration demonstrates that participants’ gaze shifted toward referential targets during robot instructions, and that richer action contexts shaped both repair formulation and pre-repair visual attention. This work contributes to the dissertation by providing the methodological foundation for studying contrastive guidance and repair as temporally unfolding, situated processes in which language–vision interaction can be examined online within a dialogical setup.

Chapter 2

Language as a contrastive mechanism

2.1 Does language foreground relations?

At this stage, it should be clear that language can *structure relational perception*. As argued in the previous chapter, contrasts emerge from relations that isolated attributes alone cannot exhibit. When a linguistic description pushes relational aspects into the foreground by reasoning for, how a *Figure* stands with respect to a *Ground*, which path it traces against the established history, and what event it accomplishes, the contrastiveness becomes available as an interpretive stance. Conversely, when descriptions dwell on isolated attributes, the scene feels less contrastive because the relevant relations are taken for granted. This perspective turns a set of familiar questions into a single investigation: How does language help to foreground relations? Which *linguistic constituents* carry these relations in ordinary sentences? How are *nonlinguistic* cues such as visual context framed by the linguistic choices that we make? Throughout this chapter, I will try to address some of these questions. Although we have emphasized dimensions to breakdown relational perception on distinct levels, here I will use *event* as an umbrella term for the dimensions on which language can operate to guide and regulate relational prominence.

There is no surprise that the influence of language on event cognition has a long history. For example, early studies on motion-event cognition associated with Slobin and Talmy showed that languages routinely conceptualize two action elements—PATH and MANNER (Slobin, 1996; Talmy, 1985a). Language was also shown to distribute action elements across verbs and satellites in systematic ways (often called *verb framing*). In satellite-framed systems such as English or German, the verb readily carries Manner while Path is expressed in a particle or preposition, for example, in an expression like: *Sie schwamm über den Fluss hinüber* (Translated: She swam across the river), the manner (Swim) and path (across) are expressed in distinct constituents. Conversely, in verb-framed systems such as Spanish or French, Path is often encapsulated within the verb, where the Manner is only added if required, for example, in an expression like: *El niño entró a la casa* (Translated: The child entered the house [running]), the path (enter) is expressed in the verb.

Crucially for our purpose, we see this linguistic conceptualization of motion events not merely as a verb-framing processes, but as a core mechanism through which the syntactic preferences highlight the Figure’s relation to the Ground, i.e., *relational prominence*. Seen from this perspective, language frames the event in a figure-centric way where the event path becomes the natural attentional anchor. To support this, evidence has been presented in the literature that the path receives more attention than other event constituents forming an inherent part of the goal where the path terminates (Lakusta & Landau, 2005; Regier & Zheng, 2007). This also aligns with the general observation that, in *normative* situations, the Figure is foregrounded and relational structure often remains in the background until the description foregrounds it. Hence, when language explicitly articulates the relation of the figure to the ground rather than just highlighting the figure’s attributes, the relational prominence should be higher.

For example, consider two ordinary sentences. In one, a boundary crossing is named through verb framing such that the path is expressed in the verb, e.g., *The cyclist entered the plaza*. In the other, the verb only encapsulates the manner and the path is additionally expressed through a satellite. In this case, only one location is described: *The cyclist rode in the plaza*. Whereas the first sentence profiles a discrete change in the Figure–Ground relation (outside $\rightarrow$ inside) and invites the reader to align the Figure with a Ground boundary, the second maintains an activity within a region and leaves that boundary in the background (Talmy, 2000).

The same pattern recurs with satellites. For example an expression like: *The stroller moved onto the stairs* versus *The stroller stopped at the stairs*. The choice between *onto* and *at* subtly regulates how contrastive the scene feels, even when the physical layout is unchanged. Meaning, with respect to motion events, when a change in background is tracked along with the figure’s course of path (as in “onto”), the relations become more prominent than only when figure’s state is described (as in “at”). Similarly, aspect and viewpoint exert a similar perceptual experience. For example, a perfective form such as *The child climbed the ladder* commits to an achieved endpoint and thus sharpens the Figure–Ground alignment at the top of the ladder, whereas *The child was climbing the ladder* keeps the process in view and withholds commitment about the goal (Madden & Zwaan, 2003). In this case, the processuality of events affects the relational prominence. A change of state caused by resultatives and causatives is even more explicit such as in expression: *She pushed the door open* builds the result state into the clause and thereby highlights the boundary that matters for the task.

Crucially, all these events, when articulated, create a change in the state of the figure with respect to the ground, while keeping the ground in focus to different degrees along with the figure, which almost always remains in attention (Talmy, 1985a). One thing to note here is that, when the change in the figure with respect to the ground is explicit, it is perceived as foregrounding relational aspects more than when the figure’s change through the background is not made evident. Meaning, when the change in

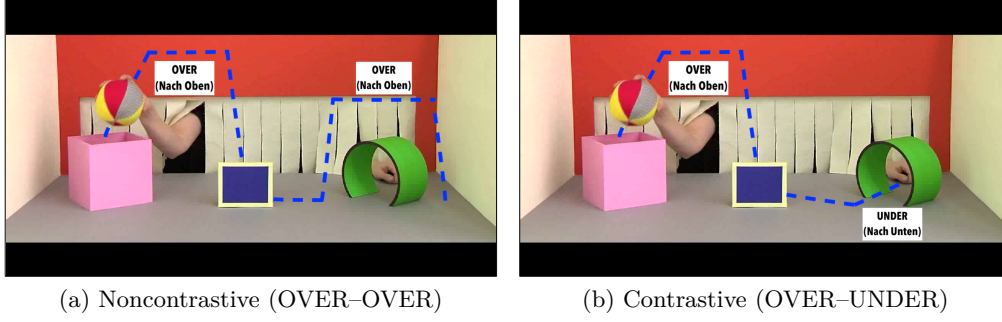


Figure 2.1: Stimuli depicting the action conditions. The blue dashed line indicates the trajectory of the ball relative to the objects. Panel (a) shows the Noncontrastive (OVER-OVER), where the action is repeated. Panel (b) shows the Contrastive (OVER-UNDER), where the second action creates a spatial contrast against the first. Adapted from Singh and Rohlfing (2025a).

ground is made more explicit with respect to the figure’s motion (or path), then the relational perception should be higher. Following this, we tested this assumption in our first study where the ground change was made more explicit with respect to the figure’s motion. If the ground change is explicit, the language marker making such relationships prominent should help to perceive events as holistic. Although most studies in the past tested change in event ground as sentential expressions through path satellites, as we have noted in the above examples, we tested event conceptualization directly through action observation visually and presented it with verbal guidance. As reported in Singh and Rohlfing (2025a), a verbal guidance (negation) bringing the past subevent (ground) into attention while keeping the current trajectory of the figure into focus was found to support the recall of the whole event. Essentially, the grounds formed against the moving figure were contrastive in nature (dimension:locative), suggesting a change in location across two subevents, as shown in Figure 2.1. The figure, being the principal entity, drew maximum attention, leading to attention toward the goal; however, when negation is used to bring attention back to the previous event (ground), the perception of action was enhanced, suggesting that when figure and ground both are brought into attention through some verbal means in a visually contrastive context, the perception of those actions is facilitated. Although the study showed contrastive effects along the locative dimension, the possibility of such effects of language intervention cannot be ruled out for contrasts at other dimensions, such as situative, functional, and so on.

2.2 Linguistic contrasts steering attention

Summarizing the first set of findings from Singh and Rohlfing (2025a), we derive three principles useful for contrastive guidance for our framework. These findings are ex-

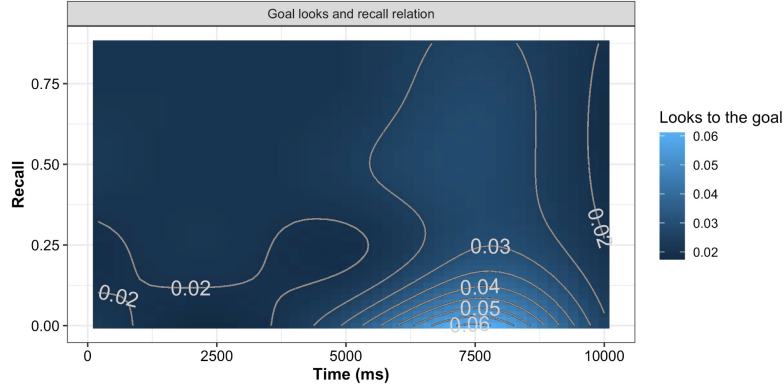


Figure 2.2: Relationship between recall score and goal-attention over time. The heatmap illustrates the probability of looks towards the goal (indicated by the color gradient and contour lines) as a function of time (x-axis) and recall score (y-axis). Lighter regions with white shading indicate a higher probability of goal-looks, showing how attention to the goal during the event correlates with later recall performance. Adapted from Singh and Rohlfing (2025a).

plicated in terms of the possibilities that language might use to structure events for relational prominence (Singh & Rohlfing, 2025a).

Past event acting as ground: Regardless of modality, the current subevent associated with the figure is always interpreted in light of what has already happened in the past (Zacks & Tversky, 2001). Hence, the past event can effectively work as a ground forming experience. Our first study exploited two forms of relations when compared from our dimensions as illustrated in Figure 1.3. Firstly, at the *locative* dimension, the change in the path position was manipulated in opposite directions to make the relations perceptually contrastive at the locative dimension (UP vs DOWN) (Singh & Rohlfing, 2025a). Here, a plausible alternate was always available, which made the scene contrastive (cf. section 1.2.3). On top of this, a possibility of *eventive* contrast cannot be ruled out as well, such as when the current event is different or similar to the past event, making the ground (past event) more or less salient. At the linguistic level, verbal instructions were introduced to guide the motion events. Given the possibilities of language influencing non-linguistic perception in various ways, we identified that language meddled with ongoing event conceptualization, such that when the later location was negated (e.g., Not Over), it brought the ground (previous event) and figure into a single interpretative framework leading to a holistic perception for the events when they are visually contrastive (Singh & Rohlfing, 2025a).

Goal prominence: The second principle take-away from the first study was regarding goal prominence. A robust finding in event cognition literature in general and motion event perception in specific is that in most dynamic scenes the goal loca-

tion is attended to more than any other constituent in the entire event (Papafragou, 2010; Zacks & Tversky, 2001). A reason for this behavior can be provided through evolutionary as well as cognitive perspectives. A general consensus is that the attentional system is goal-driven and hence it is a cognitive general propensity that guides our attention towards the goal in linguistic as well as non-linguistic tasks (Austin & Vancouver, 1996). Language in this case spotlights goals which can be modulated through specific verbal means (Singh & Rohlfing, 2025a). In the study, we suggested two ways to achieve this: a) spotlighting the current subevent through presenting verbal instruction before the specific subevent to spotlight that particular subevent more than the goal; and b) using negation to direct attention toward the prior event for locative contrasts (e.g., Now Over » NOT Over). While the former strategy debiases the attention from the goal, negations scaffold the current event in light of the past and alternate possible events, which are essentially forming “grounds” and bringing those grounds into attention and highlighting relations rather than mere current (isolated) subaction. Since in the literature an anticipatory attention to the goal (goal-biasness) was also found to cause coarse perception of mid-events (Papafragou, 2010), and in turn was detrimental for memory of the fine-grained events which happened before, there seems to be a general trade-off between attention to the goal and memory of the previous subevents. This phenomenon was explicitly tested in Singh and Rohlfing (2025a). Across two eye-tracking studies, it was found that when the attention to the goal was highest, the recall for the preceding subevents was lower (Figure 2.2) (Singh & Rohlfing, 2025a).

An important takeaway from this finding is that, given that figure attracts maximum attention in motion events, the goal related to the path introduces a goal-biasness. And in such cases, highlighting the previous event through contrastive verbal guidance (negation) brings the ground (past events) into attention along with the figure.

Action Packaging: Spotlighting particular aspects of the event may bring attention to the aspects in action which otherwise do not remain in the focus generally, and might be overridden by language specific properties. In this case, the figure traced path remains in the focus rather than the manner or how the object moves across the two subevents. Goal-debiasing, although seems to be not good for event conceptualization, in fact aids memory for the less salient aspects within an event. Whereas the goal-debiasing can be done through contrastive verbal guidance for visually contrastive actions, another macro-method to modulate goal-attention is verbal temporal synchrony with the action. Specifically, higher goal-biasness is observed when the instructions are overlapping with the action, however when verbal instruction is presented before the onset of action, the goal-biasness is mitigated (Singh & Rohlfing, 2025a). Meaning, when the objective is to bring attention on the fine grained aspect of the events, presenting instructions before actions was proved to be beneficial, as shown in Figure 2.3.

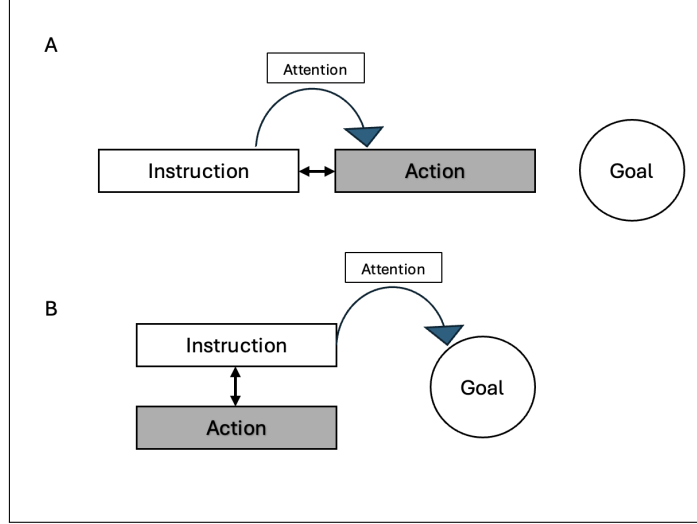


Figure 2.3: Two forms of multimodal synchrony (Singh & Rohlfing, 2025a): (A) sequential and (B) simultaneous. In the sequential presentation (Experiment 1), attention is directed toward the immediate action (path) as the instruction precedes and introduces the action. In the simultaneous presentation (Experiment 2), overlapping instruction and action are expected to shift attention directly to the action goal due to ever-present goal bias in event.

The above patterns of findings show how linguistic constituents can act as a *meddler* and *spotlight* to highlight relational aspects by making grounds salient for locative and eventive contrasts (Figure 1.3). Hence language can act as a *contrastive mechanism* alongside events to select or sharpen specific relations in the event representation. When interpreted along the contrastiveness scale, instructions in specific forms may move the contrastive perception along a *contrastiveness scale* whose primary axis is figure (path to goal) and secondary axis is background, defined as previous event (for eventive) and plausible location (for locative). Thus, our study identifies two forms of contrast emerging at linguistic level where grounds were formed at two dimensions: eventive (previous events) and locative (alternate possible locations afforded by the scene).

Linguistically constrained contrasts in motion events Though in our study, we have explicitly guided motion events by aligning linguistic constituents with action demonstrations – and hence multimodal – syntactic constructions themselves can also cause relational prominence to emerge through different linguistic mechanisms. Although the cognitive effects of different syntactic constructions have already been comprehensively dealt with in Talmy’s seminal work (Talmy, 1975), the notion of contrast and how it may emerge at different scales is nonetheless a novel perspective and provides a way to formalize those perceptions for practicality at distinct linguistic conceptualizations. Here, we see that each component of the action as defined in Singh and Rohlfing (2025a) – Figure, Path, and Manner – can work together to form different degrees of

contrast. Each component, when handled in a specific way under syntactic construction, seems to be eliciting contrastive perception. Hence, below I attempt to show a few examples of how the contrastive scale as shown in Figure 1.2 might reflect different linguistic constructions framing figure and ground and simultaneously highlighting relational prominence to manipulate contrastiveness. Crucial for this perspective are the approaches that can be ascribed to the phenomenon to explain these effects. So far we have considered possibilities of language acting as meddler and spotlighting; we may observe that along with these explanations, other possibilities might play a crucial role in defining some of the reasons behind how language might create an impression for contrasts. In light of the experiment and the components established in the literature, such as Figure, Path-goal, and Manner, here we try to find out the way to locate these constituents on the contrastiveness scale when they are expressed in linguistic constructions.

Figure For figures being the locus of attention in most events, the figure’s existence against the background can be taken as highlighting the contrast at different levels. For example, when an obstruction is introduced between a figure and ground, the contrastiveness is perceived to be the highest. With regard to the role of language influencing event perception, it was already registered that the figure being the principal entity remains in the spotlight of attention due to biases acquired through the frequent use of figure-centric constructions in a language (Talmy, 2000). And when the usual flow of the figure with respect to the ground is obstructed, the relational element is foregrounded (as in example 1). In a usual case, when there is no obstruction, the flow of the figure is lucid with maximum attention on the figure (as in example 3), where one hardly bothers about the background. A few sets of example situations describing differences in contrasts linguistically are provided below:

- (1) High Contrast: The bench **blocks** the crosswalk. (obstruction)

In the above example (1), the obstruction between figure and ground gives rise to contrastive perception. Similarly, passive situatedness gives rise to different situative contrast. In this case, the contrast is not as high as “obstruction”; nevertheless, it exhibits a sense of contrast through passive situatedness as can be read in example (2):

- (2) Mild Contrast: The bench **stands on** the crosswalk. (passive situated)

When the figure stands in proximity to the ground without obstruction, low contrast is perceived, such as in sentence (3):

- (3) Low Contrast: The bench is **near** the crosswalk. (proximity)

Path goal As it was suggested in Singh and Rohlfing (2023), the figure’s movement against the ground in a specific direction leads to goal biasness where the goal is prioritized within the locus of attention. This is not peculiar to motion events where the figure is traversing toward the action end. In the past, multiple studies in adults as well as in children verify this finding, suggesting that language and general cognition prioritize and spotlight the event goals along which the path terminates (Lakusta & Landau, 2005; Regier & Zheng, 2007). Crucially, the change of the figure’s position with respect to the ground gives rise to differential perception of the contrastiveness. For example in example (3), the movement of the stroller (figure) along the ground (stairs) shifts the perceptual trajectory from the figure being in a different ground to a *new ground*.

- (4) High: The stroller **has moved onto** the stairs. (result achieved)

In the above example, the perceptual change of ground tied to a figure comes out more explicit than in examples (5) and (6). In an obvious way, this change can be made more explicit by providing the two grounds against which the figure traverses. The language in this case seems to be spotlighting each ground one after another, such as in the sentence, “the stroller has moved onto the *stairs* from the *floor*”. It is crucial here to also notice that providing alternate grounds (i.e., possible locations) as defined under *locative contrasts* increases the likelihood that the contrast emerges; this aligns well with why locative contrast also relies upon implicit possible locations.

- (5) Medium: The stroller **was moving onto** the stairs. (incomplete)

Similarly, when the figure traverses with the change state “yet to be achieved”, the ground does not conform to a transition, hence perceived as mildly contrastive as exemplified in example (5) (Madden & Zwaan, 2003). Furthermore, when the figure is perceived to be within a specific ground without depicting a transition as in high contrastiveness or depicting the change in state as “yet to be achieved”, it is perceived as a figure lying on ground with no high relational prominence between figure and ground as in example (6).

- (6) Low: The stroller **was on the** stairs. (locative reading)

Manner Another way to depict the change of state is through the manner of action. In satellite-framed languages, manners are expressed through verbs in syntactic constructions (Slobin, 1996). Crucially, the way manner might depict the end state can be subtle or explicit. To take an example from a study by Kaup (1997), it was shown that the way (manner) of *end state* achieved by verbs has an effect on the accessibility of the object under consideration. For example, a verb depicting a destruction manner (e.g., burning) elicited more contrast than a verb eliciting creation (baking). This was

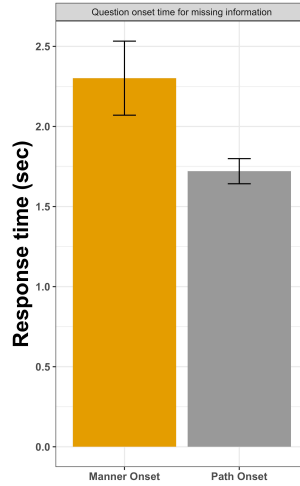


Figure 2.4: Response times for questions about missing information. The bar chart displays the mean response time (in seconds) for questions targeting the manner of the action (orange bar) versus the path of the action (gray bar). The results indicate that path-related questions were answered significantly faster than manner-related questions. Adapted from Singh and Rohlfing (2024)

further supplemented by the complementary effect of contrast elicited through linguistic negation, such as a negated object, although usually more difficult to access than asserted objects, in cases where manner was tagged with destruction, the accessibility difference was eliminated, such as in sentences like:

- (7) “Elizabeth tidied up her drawers. **She *burned* the old letters but not the photographs.**” (Example taken from Kaup (1997))

As suggested in Singh and Rohlfing (2025a), eventive and linguistic contrasts are not working against but to complement each other; provided a context, the type of verb specifying the way the end state of an entity is achieved shall produce an impression of difference in contrast. The complimentary effect of linguistic negation with the type of verb (destruction) on object accessibility, as in example (7) might suggest that the verb manner does create a different degree of contrast leading to the processing cost alleviation associated with negation. Suitable examples for such cases are compared in examples (8) and (9). Whereas in example (8) the principal entity is destroyed, where the change of state is explicitly brought forward achieved against a ground, example (9) seems to only underspecify the end state, which will be perceived as less contrastive than the former.

- (8) The glass **shattered because it fell** off the table. (manner explains result)
 (9) The glass **moved** off the table. (underspecified, weaker explanatory contrast)

When we interpret these findings in parallel to our contrastive scale, the *Figure*, *Path-goal*, and *Manner* together form the primary axis of contrast, such that the descriptions that commit the Figure to a boundary crossing (likely creating an obstruction for the figure to the ground) yield the strongest contrast, adjacency terms without suggesting a change to the figure’s existence to the ground (e.g., *to/at*) exhibit mild contrast, and mere vicinity (e.g., *near*) lowers it. Thus, at the *Figure* level, contrast depends on how forcefully the Figure’s presence restructures the scene, such as an *obstruction* (the bench *blocks* the crosswalk) ranks higher than *passive situatedness* (the bench *stands on* the crosswalk), which in turn exceeds simple *proximity* (the bench is *near* the crosswalk). *Path-goal* characterizes how the figure traverses through the path and as a result of which how the ground changes and comes into attention as a product of relational prominence. Importantly, in this case, the course of change of state becomes the locus of contrast modulation rather than the figure itself. Finally, the *Manner* acts as a graded modulator which elevates contrast when it is atypical or extreme, when it causally explains the change of state (as in, *the glass shattered because it FELL*), or when attention is explicitly directed to the “how” of the motion. In other cases, manner leaves the scene closer to a low-contrast with underspecified conceptualization.

In addition to motion events, which can naturally serve as test-beds for contrastiveness due to their components (*Figure*, *Path-goal*, and *Manner*) eliciting dynamics that can be perceived as contrasts (Talmy, 2000), linguistic constituents can in themselves elicit contrast on a scale (Langacker, 1987). A selection of these linguistic contrasts and their realizations are shown in Table 2.1.

2.3 Implicit and explicit contrasts

As we have seen, at a relational level, the degree of contrastiveness depends upon how the different or opposing elements are conveyed in language for the provided dimension. These constructions could concern the relational dynamics between *Figure* and *Ground* that allow us to observe contrast on different levels. Depending upon the dimensions, as suggested in Figure 1.3: *Situative*, *Functional*, *Locative*, *Eventive*, and *Social* – the realization of contrast can be stated implicitly or explicitly. In general terms, implicit and explicit contrasts differ primarily in how the opposing or differing elements are communicated (Sperber & Wilson, 1986).

On the one hand, explicit contrast directly states or clearly expresses, often using overt markers like “not” or clear negations (e.g., “This is red, not green”) to some properties related to the figure or contrasting with experiential ground (Horn, 1989; Wason, 1965). On the other hand, implicit contrast conveys meaning indirectly or through implication, where the contrast is understood from context without being stated outright (e.g., using different words for two events without explicitly negating one) (Au & Laframboise, 1990; Roseberry et al., 2009). Crucially, the main function of both contrasts is to

Constituent	Example contrast	Effect on event contrast
(i) Verbs	<i>enter</i> the plaza vs. <i>ride in</i> the plaza	Boundary-crossing verbs assert a discrete change in the Figure–Ground relation (Talmy, 1985b).
(ii) Satellites	<i>stop at</i> the stairs vs. <i>move onto</i> the stair	Path markers toggle contrastiveness by encoding vicinity vs. boundary-crossing (Talmy, 2000).
(iii) Resultatives	<i>push the door open</i> vs. <i>push the door</i>	Resultatives adding change of state, sharpen the contrast between initial and final configurations.
(iv) Aspect	<i>was climbing</i> the ladder vs. <i>climbed</i> the ladder	Perfective event foregrounds completion (contrastive); imperfective foregrounds ongoing process, backgrounds endpoint (Madden & Zwaan, 2003).
(v) Info. structure	<i>rolled INTO the box</i> vs. <i>It was into the box that ...</i>	Focus particles, clefts (<i>only, even, just</i>) create contrast sets over possible Grounds, foregrounding one option against alternatives (Rooth, 1992).
(vi) Negation	<i>moved</i> above, <i>not through</i> it	Negation creates explicit contrast, between realized and non-realized events (Köller, 2016) and actions e.g., path and manner (Singh & Rohlfing, 2025a).
(vii) Comparative	<i>drifted closer to</i> the halt; <i>farther</i> than the balloon	Scalars grade distance along a Path toward or beyond a Ground, aligning with locative contrast.
(viii) Deixis	<i>this</i> path vs. <i>that</i> path	Deictic anchors Figure–Ground alignment relative to hearer and intensifying relational salience.
(ix) Prosody	Pitch accent: INTO the BOX	Prosodic prominence on path or manner guides attention to relations bringing contrast without changing content.

Table 2.1: Linguistic mechanisms modulating relational prominence. Few examples of how different constituents (e.g., verbs, negation, aspect) structure the Figure–Ground relation to shift perception along the contrastiveness scale.

bring attention to the alternate through either a) inhibiting the already established link through negations in the explicit case or b) directing attention to the alternate representation, which indirectly suggests the link was not correct and the provided link is relevant in the situation (Nordmeyer & Frank, 2014).

For the given contrastive dimensions, implicit and explicit contrasts may work differently depending upon the figure and its associated ground. For situative context, implicit contrast may meddle with the ground by providing additional structuring—crucially without negating it. Such as in the scenario of left-vs-right hand driving from section 1.2.1, the agent might be corrected with respect to the ground by only a single sentence without negation, as in, “It is UK” (hence agent infers that UK has Right-Hand-Drive unlike her country). Here, two things in the situation are cleared: firstly, the implicit contrast between two countries with respect to driving rules, and secondly, essentially two different labels for the alternates which were in consideration, i.e., UK and home country. Conversely, if the aim had been to explicitly contrast the two situations, one could have used negation with respect to only one label, such as “It is *not* your home country (hence inferred, unlike my country, in this country it is

right-hand drive).” Crucially, negation here acting as a linguistic contrast provides a contrast against the target ground itself, making the difference obvious; on the other hand, implicit contrast creates entirely a new ground by introducing an alternate.

This can clearly be observed for functional contrasts. For example, consider a child observing her mother preparing dough; at one moment she asks her mother, “Are you *mixing* it?” And the mother replies, “I am *cutting* it.” The response from the mother creates an implicit contrast between two actions, such that now the other action is also implicitly decoded in the context (H. H. Clark, 1996). Had there been a case where negation was used to explicitly counteract the ground (assumed action), the mother could have responded with something like, “I am not mixing, but cutting.”

Similarly for locative contrasts, an implicit or explicit contrast can be defined in terms of alternate locations that are probable in a context. A suitable example can be taken from Singh and Rohlfing (2023), where the location of the ball moving in the opposite direction was either explicitly negated with reference to the ground (past action), such as “Not Up” for a motion that was previously “Up” but now “Down”, or implicitly contrasted through the actual movement directed using assertives such as “Now Down” given that previously the motion was “Up” but now it was “Down”. Crucially, when the present location (figure) was negated against the ground (previous location), it foregrounded the relational aspect due to the current figural event being interpreted in light of previous event, rather than introducing a new label which would be an act of implicitly contrasting the two opposite locations. This connects nicely to the eventive contrasts, as shown in Figure 1.7, where if the present subevent is negated (e.g., put this not vertically), a rich connection is created among the current (figure) and prior event(s) (ground), foregrounding the relational aspect, as compared to when a new label (e.g., put it horizontally) is used to implicitly contrast the present event. In an unfolding action, if an action sequence contains a visually contrastive pair (e.g., an “Over” move followed by an “Under” move), placing negation on the later subevent — *Now Over—Not Over* — references what has just been established, inhibits the new label, and helps encode the distinguishing prior-event structure rather than defaulting to new events (Singh & Rohlfing, 2023, 2025a). This turns us to the question; what explicit contrasts such as negations can do for actions and events that simple assertions or implicit contrast cannot?

2.4 Negation

On an odd day, I was in my basement doing laundry when I came across a statement from my landlord for all the residents. The instruction referred to a switch located in an obvious position where a light switch is typically found. The switch was supplemented by a statement: “Heizungs Notschalter! Kein Lichtschalter!!!” (English: Switch for the heater! Not for lights!!!). The exclamation marks are particularly important here,

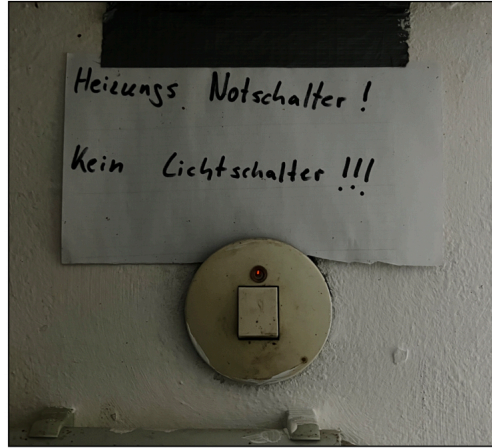


Figure 2.5: An example of negation being more impactful than assertions. The image shows a handwritten note placed above a switch, reading "Heizungs Notschalter! Kein Lichtschalter!!!" (Switch for the heater! Not for lights!!!). This illustrates how negation ("Kein...") is used to explicitly inhibit a prepotent response (turning on what looks like a light switch) by opposing the expected function with the actual one, often conveying stronger caution or emphasis than a simple assertion would.

with negation seeking more caution through three exclamation marks compared to the assertion, which has relatively one (Figure 2.5). Why would one choose such a construction? Is it pragmatically more plausible than assertions, or does negation state more than what an assertion can convey? Although negation is generally found to be more difficult to process than assertions (Just & Carpenter, 1971; Kaup & Dudschig, 2020b), why do we seem to use negations so frequently in day-to-day scenarios? In this section, we will discuss similar questions concerning negations, especially focusing on their repercussions for pragmatic principles and what they mean for actions.

It appears that negation can be a powerful device to inhibit prepotent responses, an element in our framework that we have been calling *figure* (having high goal affinity). Crucially, this inhibition is achieved by directly targeting expectations, an element in our framework that we've been calling *ground*. Negation is thus an *explicit* contrast device that asserts a present state of affairs *by opposing it to an alternative* salient in a context (Horn, 1989). A dominant notion in psycholinguistics is that negative statements generally take more time to process than their affirmative counterparts, a reason being ascribed to the alternate state of affairs being activated before the actual state is represented (Kaup & Dudschig, 2020b; Kaup et al., 2007). The notion is primarily connected to a specific cognitive processing strategy popularly known as the "two-step" processing mechanism. For instance, saying "not red" would first activate features related to red and then finally the intended meaning (Kaup et al., 2007). This

proposition is supported by early studies showing that a two-step mental simulation takes place when someone encounters a negated sentence. For example, in a study by Kaup et al. (2007), participants were shown a sentence, such as “There was no eagle in the sky,” and thereafter, a picture matching a negated state of affairs (an eagle with stretched wings) or an actual state of affairs (an eagle with folded wings). It was found that after a short Stimulus Onset Asynchrony (SOA) of 500 ms, participants responded faster to a picture matching the negated state of affairs than to a picture matching an alternate state of affairs. Crucially, this pattern was reversed when sufficient time for a response was provided, which presumably allowed participants to simulate the situation after reading the sentence. The results suggested that participants relied predominantly on mental simulation while processing negative sentences. Based on findings like this, conceptual access under negation has been shown to be governed by a two-step cognitive process: at the first step, activation is guided by the propositional structure, and at the later step, by the situational model (for a review, (Kaup & Dudschig, 2020a)).

Other theories, however, predict that the activation of the alternate state of affairs is sensitive to the context; for instance, when the positive counterpart is already prominent before the negation is used, the two-step process might not necessarily take place (Kaup & Dudschig, 2020b; Nordmeyer & Frank, 2014). Thus, a context licensing the plausible use of negation might, in a few cases, be facilitative or at least promote negation usage (Wason, 1965). Nevertheless, the effect of two-step processing is observed not only in linguistic tasks but also in nonlinguistic tasks such as actions. Past studies have shown that negation leads to action inhibition, where the reason also being cited as that alternate state of affairs being activated before the actual state of affairs. This simulation has been shown to be constitutive of domain-general cognitive processes, which directly connect to actions and embodiment accounts. Along this line, it is suggested that negation also relates to the sensorimotor system, which involves general cognitive mechanisms such as response inhibition. More specifically, in an electroencephalography (EEG) study by de Vega et al. (2016), participants were asked to read an assertive or a negative sentence while doing a NoGo task; it was found that theta oscillation – a neural marker of response inhibition – was smaller in the negative than in the assertive condition (de Vega et al., 2016). It is worth noting that, the studies in favor of the two-step cognitive process usually focus on picture recognition, inhibitory control, or sentential negation, where the accessibility of a critical concept is measured through the recording of reaction time and accuracy.

Coming back to our example as in Figure 2.5, where the required action for the switch is bound around negation and the goal of the conveyor would have been to either inhibit highly routinized responses (unintended switch presses) or to make people aware not only of instantaneous response but also of future actions. In this sense, using negation might not only connect with action on a surface level (to inhibit repetitive actions) but also at a deeper conceptual level to modulate or possibly change the response in future.

Crucially, a goal which is mostly targeted while scaffolding others (Wood et al., 1976). Nevertheless, if negation connects with action at a deeper level compared to mere assertions, it may have the potential to be used as a scaffolding mechanism, especially for actions where the user will be aware of future actions to be performed in a given context, known as enabling (Buschmeier et al., 2025). Aligning with this hypothesis, in one of our studies, we tried to scaffold users’ attention in a Human-Robot explanatory dialogue using negations for possible actions related to an object on a touchscreen (Groß et al., 2023). In this study, we found that although actions took more time to initiate when guided through negations, the actions were performed more accurately compared to when an action was guided through assertion, suggesting a controlled process during action execution for negations. Following this, we underscore the potential of negation as contrastive guidance especially when two actions compete with each other afforded by common object. An explanation of this effect can also be provided within our current framework of relational prominence (Chapter 1).

As proposed in our framework, contrast is supposed to foreground the Figure–Ground relations over isolated attributes. Here, the ground was related to the highly salient action which was expected and hence negated; against this action, the intended action (figure) was scaffolded. This fits well with our definition of functional contrast, where the expected action acting as an experiential ground (afforded action) is contrasted against the intended action. Negation as a contrast was found to achieve exactly that, as it *selects a ground* (the prior, expected, or contextually live alternative) and *profiles the figure* (the current state) *as not that*, thereby increasing relational prominence on the contrastiveness continuum. In this sense, negation is not merely a truth-functional operator; rather, it is a context generator provided that it links the current event to the prior events (Singh & Rohlfing, 2025a). As provided in the example above, a practical use of negations thus is in a scenario where the unwanted action is supposed to be inhibited such that it should not occur in the near future and at the same time provide an alternate correct action. In this sense, negation can be more powerful than assertions because it can recruit both pragmatic (back-referencing) and domain-general inhibition mechanisms to steer attention during event apprehension. Aligning with this theory, negation was found to contribute to two core mechanisms through a series of studies as it was reported in Singh and Rohlfing (2025a).

Context generation via alternatives. As suggested in Singh and Rohlfing (2025a), when negation is used in a contrastive context (locative and eventive), it leads to better action understanding through action recall (Figure 2.6). The study further highlighted that contrastive instruction differs from mere assertive instruction due to the role of negation in generating a rich contextual effect that is appropriate for the perception of contrastive actions. In a contrastive action (Over vs Under), the present subaction (Figure) was the opposite of the subaction which was previously experienced (Ground). Thus, guiding the present action through negation entails re-

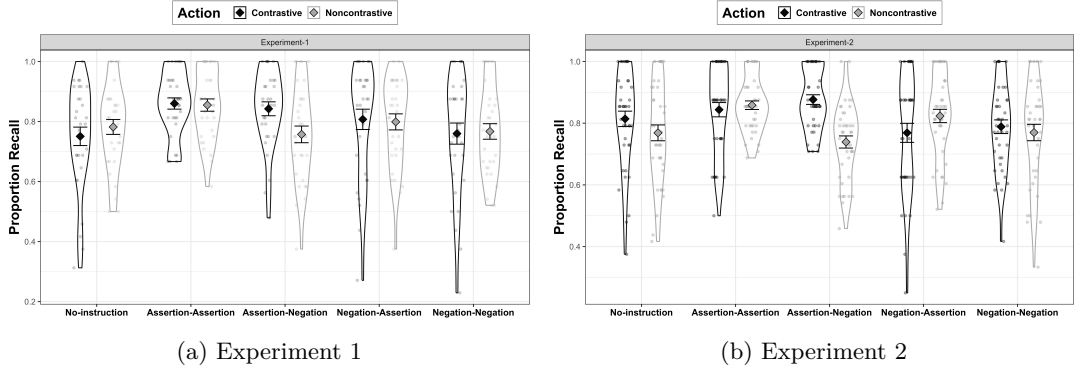


Figure 2.6: Recall for contrastive and noncontrastive motion events as reported in Singh and Rohlfing (2025a)

ferring to the Figure (present action) and Ground (previous action) simultaneously (e.g., Now *Over* $\rightarrow$ NOT *Over*). This as a result brought both Figure and Ground into a single frame of attention, and hence prioritized the *relational* aspects rather than mere attributes related to the present action associated with figure. As a result, high relational prominence offered a holistic perception of the entire event and the recall of such events was higher compared to when only the figure and its attributes were in attention. The interpretation aligns with the premise that contrastiveness depends on prominence established between Figure and Ground during apprehension, and this is maximum when the present subevent is interpreted along with the event which was previously experienced (Ground) (Section 1.2.4).

Inhibition of prepotent responses. When we negate the present subaction at issue, it implicitly suggests the alternate action, and if the alternate is already a part of the experiential ground, it is brought into attention (Nordmeyer & Frank, 2014). Hence, negation achieves two crucial roles for contrastive events: a) brings the alternate subevent (ground) into attention to generate a rich context and b) inhibits the current response through sensorimotor meddling (de Vega et al., 2016). Whereas context generation was already dealt with in the previous section, the inhibitory function is achieved simultaneously, which is at issue here. Supplemented by the findings from eye-tracking in Singh and Rohlfing (2025a), it was shown that negation, in addition to creating a rich context, also attenuates attention from the *Path-goal* as measured through fixation on the event-final goal, which otherwise remains in the focus of attention—a phenomenon dubbed as “goal-biasness” (Lakusta & Landau, 2005). A possible explanation for such inhibition was that negation, by referring to the prior subevent (Ground), resets the default locus of attention from the goal to the mid-events. This achievement is not trivial, since in most events many non-salient elements are coarsely perceived, and negation in such cases can act as a guidance instrument to elevate the perception of such elements. Crucially, in

Singh and Rohlfing (2025a) the results were discussed with the potential of negation to bring the manner of an action into attention through goal-debiasing, which forms an integral part of the *path* in addition to *manner* in an action event. Motivated by this finding, and testing this potential, our group implemented a follow-up study in Human-Robot Interaction where action manners specifically were guided through negations and found that manners were performed more accurately by the participants when guided through negations, such as *don't push, but turn!* (Groß et al., 2023). We turn to this finding when we discuss how negation can enhance contrastive perception at the functional dimension in the upcoming section.

Revisiting multimodality (action-instruction synchrony). The timing of instruction in relation to the provided action is perhaps a very crucial factor when we talk about attention on micro and macro constituents in an action ¹. Whereas macro aspects generally relate to the overarching goal of an event structure, micro aspects are more nuanced and subtle aspects. An example from Singh and Rohlfing (2025a) could be that action final goal is related to the macro and subevents before that are micro aspects. In this sense, micro aspects are tentative or subgoals which occur as a part of event structure and macro aspects are global goal and hence provide the entire event with a specific termination. Although during the event apprehension, the give and take between these aspects are taking place through a shift in attention, which was popularly known as vertical and horizontal shifts within and between events (H. H. Clark, 1996), another way of guiding attention is through tailored instructions. One aspect of such tailoring we investigated in Singh and Rohlfing (2025a) using instruction-action synchrony. Motivated by the theory of Intersensory Redundancy Hypothesis (IRH) and acoustic packaging (Bahrick & Lickliter, 2000), the study tested two ways of guiding actions: a) instruction before action and b) instruction along action (Figure 2.3). When instructions precede a subaction, they function as an ostensive cue, such that they announce that something relevant is about to “happen” and thereby draw sustained attention to the immediate subevent rather than to the action final goal (Singh & Rohlfing, 2025a). In this sequential presentation, it was found that language temporarily decouples the observer from the default goal-bias of the sensorimotor system and keeps the fine-grained subevents (the current move of the Figure) in focus, as shown in Figure 2.7 (Singh & Rohlfing, 2025a). We interpreted this findings so that presenting an assertive-negative sequence before the corresponding subaction allowed participants to link the later negated subaction back to the previously affirmed one, building a richer context in which Figure and Ground (current and prior subevent) are processed together (Singh & Rohlfing, 2025a). Conversely, when instructions overlap

¹The distinction between micro and macro aspect of the action relates to how the action goal and subgoals are apprehended within an action. For instance, while perception at the micro level specifies the current move under consideration, the macro structures are global goal and sit higher on the hierarchical event structure.

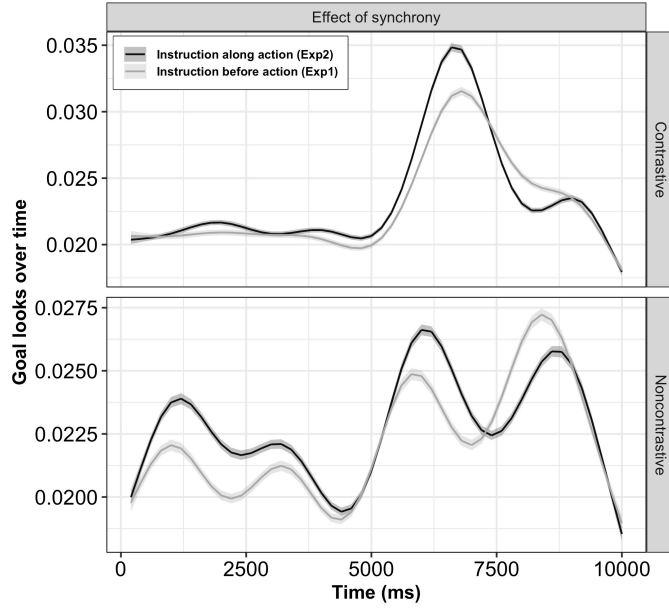


Figure 2.7: Effect of multimodal synchrony on goal-attention as reported in Singh & Rohlfing 2025: Comparison of goal-looks over time for instructions provided before the action (Exp1) versus instructions provided along with the action (Exp2), across both Contrastive (top panel) and Non-contrastive (bottom panel) contexts.

with the unfolding action, the instructions “packaged” the current action in a specific way, such that a temporal co-occurrence of verbal and visual information yields intersensory redundancy, which tends to be perceived as a single unitary event and, by default, re-privileges the goal over the mid-event details (Bahrick & Lickliter, 2000; Bahrick et al., 2002). We interpreted this result as: In simultaneous presentation, language no longer primarily spotlights an upcoming subaction but rather “rides along” with the sensorimotor routine, and the cognitive system tunes itself into its learned habit of goal-first processing. As a result, the guidance becomes more tailored, such that overlapping assertive instructions best support noncontrastive actions, whereas overlapping assertive-negative instructions selectively enhance the encoding of contrastive actions by reducing goal looks and promoting a more holistic perception of the event structure (Singh & Rohlfing, 2025a).

2.4.1 Negation across contrastive dimensions

Having laid down core findings from our studies, it is fruitful that we see how explicit contrast, such as negation, extends its principles to all the contrastive dimensions that we have enumerated in Section 1.2. Although our past studies have largely incorporated and tested three forms of contrasts along these dimensions—**Locative and Eventive** (such as in Singh and Rohlfing (2025a), Singh et al. (2025)) and **Functional** (such

as in Groß et al. (2023))—other dimensions of contrast can also be effectively guided through negations. In this section, I supplement findings for all the listed dimensions of contrast with examples of how negation can be effectively used as verbal guidance for enhancing contrastive perception. Here we will see that when the scene provides an opportunity with alternates relating to contrastive principles, the contrastive properties of negation integrate naturally with those dimensions.

Situative: As suggested in Section 1.2, situative contrastiveness in a scene depends on the extent to which relational aspects (such as the Ground) are foregrounded over the Figure or its attributes. Most situations in day-to-day scenarios foreground the Figure and its attributes unless our expectations are violated against the experienced Ground, for which relations become more pronounced. When the situation provides such an opportunity, our cognition does not fail to bring relations into the foreground. We will see that, on the one hand, such scenarios are well described through negations; on the other hand, this has repercussions for general cognitive processes. Situations eliciting contrast, such as the one portrayed in Figure 1.1b, naturally activate negation as a plausible candidate, such as:

(10) The bin is NOT at the side, but in the middle.

This construction might not be plausible for normative situations, such as the situation depicted in Figure 1.1a. Hence, negation would be used only if there is a ground against which the actual state of affairs is contrasted. Crucially, using negation in this case makes the ground at the side more prominent, and negation provides a pragmatic way of achieving this prominence (Wason, 1965). By negating the figure in a particular situation, it implicitly recruits a background of experiences as a ground against which the current situation is to be read. In other words, saying that some state of affairs is *not* the case presupposes that a different, often more normative, state of affairs is live in the background, and this experiential ground becomes more salient once the figure is explicitly contrasted against it (Nordmeyer & Frank, 2014). Hence, an utterance such as *The bin is NOT at the side* invokes a normative placement as ground (which is an expectation) and foregrounds the odd relation (bin in the middle), which moves the scene rightward on the contrastive scale (normative $\rightarrow$ implausible). In the situative dimension, the “side” location is part of a backgrounded script for where bins “belong” in everyday streetscapes, which typically remains implicit and therefore unnoticed. Negation thus both activates this normative alternative and marks the actual state of affairs as a deviation from it. Hence, rather than merely reporting where the bin is, the utterance invites the observer to consider why it is *not* in the routine place and what this might imply (e.g., something is wrong with the canonical routine, someone has moved it). In doing so, attention is shifted away from the bin’s intrinsic properties and toward the figure–ground relation and reasoning for the figure’s existence against that particular

ground that now carries the contrastiveness, i.e., the misalignment between what the situation *is* and what, given prior experience, it *ought* to be .

Functional: A concrete implementation of functional contrast comes from one of our studies (Groß et al., 2023). In this study, we used participants’ experiences with day-to-day objects. Based on our definition, a situative contrast would need a *Ground* and *Figure* standing in contrast to each other. Hence, the Figure in this paradigm would be based on how someone would typically act upon certain objects; for instance, a button on any device would be pressed, however, there is an alternate possibility that such a button also affords to be turned (Gibson, 1979; Norman, 1988). Recalling our ideas from previous studies, negating the Figure against the Ground knowledge can better support the event conceptualization for manner learning (Singh & Rohlfing, 2025a). Thus, we negated the prepotent manner and provided the alternate manner in an instruction such as “*Don’t press—turn!*”. Here, the guidance explicitly contrasts the intended manner (Ground) with the manner driven by prior expectations (Figure) about the object. In other words, the participant’s first impulse (to press) is treated as the Figure, and the instruction introduces an alternative, less routine action (to turn) as the new Ground. The study showed that negating the biased action and immediately asserting the correct one led to more accurate action performance (as measured on the touch screen). From the present perspective, this means that negating a manner that mismatches the tool–goal schema inhibits a well-practiced routine and highlights a functionally appropriate alternative (R. Cooper & Shallice, 2000). The underlying figure is the practiced affordance relation between tool and target action (e.g., “this kind of button should be pressed”) (Gibson, 1979), and the negation explicitly cancels this prepotent routine while providing a corrective continuation (“*not* pressed, but turned”). By suppressing attention on the prepotent manner and immediately offering a compatible one, the utterance pushes the learner to attend to a less frequent, but more appropriate, way of acting (de Vega et al., 2016). In terms of our contrastiveness scale, the increase in functional contrastiveness can be visualized on the manner dimension, such that when there is a competing manner for an object, the functional contrast emerges; in this case, using negation to inhibit the frequent manner and bringing attention to the ground through asserting the correct manner leads to better action understanding. A reason for this effect could also be ascribed to high relational prominence between the two possible actions, where a holistic understanding of both actions is achieved in a functional situation. Whether such guidance can also affect future task awareness is an open question, which we are trying to understand in relation to negation’s role in action scaffolding.

Locative: An apt utilization of negation for locative contrast is realized in Singh and Rohlfing (2025a). According to our definition, locative contrasts are defined by the possible locations brought forward by the context against which the figure

can be interpreted (Herskovits, 1986). Our study fits well with this notion, where a motion event with two subsequent path trajectories (Up or Down) was standing in contrast to each other in terms of its location (Figure 2.1b). Crucially, these locations were the only options that the participant could hold as their ground (expectations), which were opposites and alternates. Hence, using ‘*Not Under*’ as verbal instruction negates the later ball location with respect to the previously established subevent, so that the current move is read through the ground furnished by the past event. As opposed to a situation where the second subevent was guided with an assertion like ‘*Now Under*,’ where the locative change was labelled afresh, saying “*Not Over*” for the same Down movement explicitly contrasted the earlier location (ground) and folds both positions into a single representational ground. Through the contrastive guidance, rather than simply naming where the ball ends up (as in the assertive case), the guidance foregrounds the relational aspects in location across subevents (previously over, now no longer over), thereby increasing locative contrastiveness and bringing the figure-ground relation (past vs. present path position) into focus.

Eventive: Similar to the locative case, an exemplary use of negation for eventive contrast appears in Singh and Rohlfing (2025a), and in part within a Human-Robot Interaction study (Singh et al., 2025). Whereas locative contrasts are spatially characterized, eventive contrasts are spatiotemporally defined, where all past events act as a Ground for the current event (Figure) which is currently attended to. In Singh et al. (2025), a sequence of actions established a context against which the current event was guided. These events consisted of a sequence of manners, such as placing a screw in a specific orientation one by one. A sequence of manners established a context against which the current manner associated with the final screw had to be placed. As depicted in Figure 1.7, a repetitive manner established a ground, and if the upcoming manner was opposite to what had happened in the past, a contrast emerged. This is similar to expressing, “previously all the screws were placed vertically, but this time it is NOT vertical”. Crucially, when a negation was used to guide the present manner while simultaneously referring back to the already performed manners, the relational prominence between the Figure (current event) and Ground (all past events) was foregrounded. We found that negating an already established repetitive action to guide a current action with an opposite manner reduced cognitive load, as shown by lower pupil dilation compared to assertive instructions (Figure 2.8). Interestingly, the effect of negation switched when used for a non-contrastive context, such as where the past manners were not repeating. Here, assertive instructions were more useful given the non-contrastive context. Thus, although negation is generally reported to elicit processing load, under a specific context (here: eventive ground), the cognitive load was lower than for assertives (Figure 2.8). We will shed more light on context in the next chapter. Here, our aim was to show that negation can bring Figure and Ground into a common frame of interpretation within a

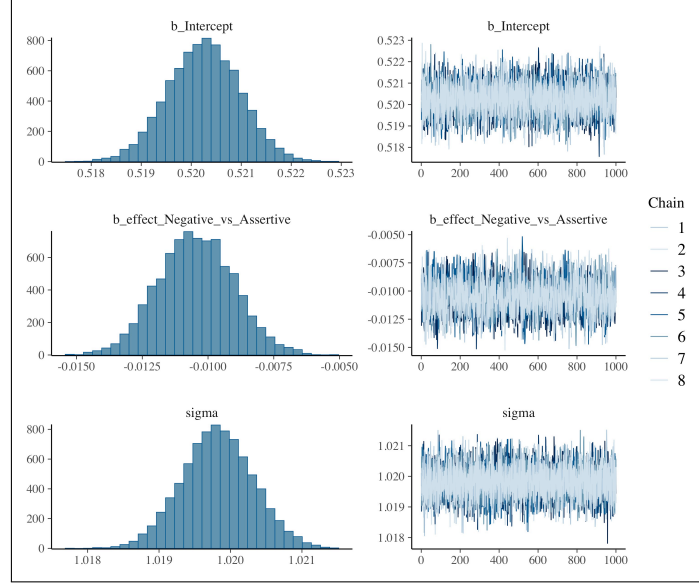


Figure 2.8: Posterior distributions for contrastive instruction on cognitive load (pupil dilation) in the eventive contrast. Negative distribution (middle row) indicates participants’ lower pupil dilation (reduced cognitive load) when processing Negation compared to Assertion after a repetitive action history.

contrastive event.

Another example of negation being used in an eventive contrast can be observed in Singh and Rohlfing (2025a). Here, the guidance pattern followed and aligned with the contrastive action sequences (e.g., *Now Over*—*NOT Over*), where eventive contrasts are defined not by a static location but by how the present subevent (figure) is interpreted against the immediate history of the event (ground) (Zacks & Tversky, 2001). In our stimuli, the ball followed two subsequent paths (e.g., *Over*—*Under* or *Under*—*Over*), so that the first subevent provided a strong expectation and hence a natural ground for the second. When both subevents were merely labelled assertively (*Now Over*—*Now Under*), the sequence remained largely a case of visual contrast; the two paths differed, but the second subevent could still be encoded as a new subevent directed toward the goal. By contrast, using an assertive–negative pair (*Now Over*—*Not Over*) for the same contrastive trajectories explicitly tied the later subevent back to the earlier one. Negating the current subaction entails referring to the prior subaction as ground; in doing so, it brings the figure and ground into a single interpretative frame. Our eye-tracking and recall data showed that this form of contrastive guidance reliably facilitated action understanding across two experiments. In other words, when the present subevent is scaffolded through negation against its own history, eventive contrastiveness increases; i.e., the sequence is perceived more as a structured whole, thereby foregrounding relational aspects.

Social: In the social domain, an intuitive case for how negation can realize social

contrast can be illustrated with an everyday restaurant script (Schank & Abelson, 1977). As argued in Section 1.2.5, social contrast arises when an ongoing state of affairs involving agents is read against a normative background of shared scripts where the script furnishes the ground (e.g., menus $\rightarrow$ order $\rightarrow$ eat $\rightarrow$ pay), and the present social move is interpreted as the figure against this background. In this setting, when the expected move goes against the script, the use of negation becomes plausible (Wason, 1965). For example, when the bill arrives before the food, an utterance like *Don't pay yet* indexes a script step out of order and therefore explicitly contrasts the current move within the normative sequence. The attempted payment is a figure that would normally be licensed only once eating is complete; hence, by negating this move, the speaker both inhibits the premature step and simultaneously evokes the proper ordering as a live alternative (Nordmeyer & Frank, 2014). In this sense, the negation does more than negate a single act; it momentarily foregrounds sequence relations (who is supposed to do what, and when) over the mere content of the act (paying). Put differently, the contrast is not pay vs. do-not-pay in the abstract, but “the attempted payment vs. the position it should occupy in the script.” With regard to our scale, such a deviation pushes the configuration rightward on the social contrast scale where the deviation from the script becomes visible as a relation, and the normative alternate is advertised in the background. Even though we did not report a dedicated study for social contrast here, the figure-ground logic seems applicable as in the other dimensions where negation can be used to flag a misaligned social move, re-align interlocutors with a shared routine, and thereby make the underlying relational structure (roles, turn-taking, sequence) more salient than the isolated action scripts (Garfinkel, 1967).

2.4.2 Formalization of contrastiveness scale

In earlier sections, I treated relational prominence R as an informal Figure-Ground distance on the contrastiveness scale. Here I attempt to make this more explicit by connecting it to a simple ACT-R style model (J. R. Anderson et al., 2004). One could ask, why an ACT-R model? ACT-R, standing for Adaptive Control of Thought-Rational, is a cognitive architecture which tries to model human cognitive processes assuming a set of core mechanisms operating for specific purposes (J. R. Anderson, 1990). One of the advantages of such models is that they are interpretable, having psychologically explained variables with a clear mapping to theory (Lebiere & Anderson, 1998). ACT-R provides us a suitable way to model our assumptions since: a) we can define each contrastive dimension and create their chunks with relevant slots for each dimension d as: *Situative*, *Functional*, *Locative*, *Eventive*, and *Social*; b) hard code the distance function between the Figure (F) and Ground (G); and c) estimate the availability of alternates to calculate the relational prominence. We now go through each of the steps to understand how relational prominence R will be calculated.

For any provided *Context* let's say (C), the *Figure* (F) will be encoded first for a particular dimension $F_d(C)$, which will be compared and interpreted against the available *Ground* $G_d(C)$ to get the distance (dissimilarity) between Figure and Ground $\Delta_d(C)$ within that particular context. Importantly, this is in line with our theoretical definition of contrastiveness as exemplified in Chapter 1. The dissimilarity between Figure and Ground $\Delta_d(C)$ is treated similarly as conceptualized in Figure 1.2. Here we formalize it as:

$$\Delta_d(C) = 1 - \text{sim}(F_d(C), G_d(C)), \quad (2.11)$$

In this equation, whereas $F_d(C)$ encodes the current relation of the Figure as observed in context C at dimension d , the $G_d(C)$ is retrieved from the Declarative Memory in the model by picking up the most appropriate “normative” configuration for this context (J. R. Anderson et al., 2004). This is done through searching for the most normative Ground for dimension d that fits the context C, as retrieved from memory $G_d(C)$, which is achieved by:

$$G_d(C) = \arg \max_{t \in T_d} A(t \mid \text{cue}_d(S)), \quad (2.12)$$

The above equation says that $G_d(C)$ is the best-matching template for dimension d , given a cue derived from the current context (C). Following this, when both $F_d(C)$ and $G_d(C)$ are obtained, a comparison between them is done to calculate the dissimilarity $\Delta_d(C)$. To compare the distance, we use partial matching (sim), and after normalizing the range $[0, 1]$ we subtract the score from 1 (perfect matching scenario) to get the dissimilarity (Equation 2.11) (Lebiere & Anderson, 1998). Though the similarity calculation method could be extended to more advanced algorithms, at this point I keep this area open for future investigation. With reference to the dissimilarity equation $\Delta_d(S)$, an intuition could be that when the value approaches 0 the distance is low and hence fits the experiential ground (low contrastiveness: right on the scale), on the other end as the value approaches to 1, the distance is maximum (high contrastiveness: left on the scale), as depicted in Figure 1.2.

In addition to the dissimilarity, another factor which was highlighted by the experiment is the *availability of the alternate*, on which case a contrastive guidance will depend. We define it in the model in availability terms, such as:

$$A_d(C) \in [0, 1] \quad (2.13)$$

$$A_d(C) = \begin{cases} 0 & \text{no live alternative,} \\ \approx 1 & \text{a clear alternative is given or retrievable} \end{cases}$$

Here $A_d(C)$ checks whether a clear alternate is present in the Ground (G), such as

OVER–UNDER, alternate manners (Vertical–Horizontal) on a specific dimension provided a context. After all these values are estimated the relational prominence $\mathbf{R}$ in a given context for specific dimension is $R_d(C)$. Hence relational prominence on dimension d is high only when the realized context is far from the experiential ground $\Delta_d(C)$ and a credible alternative is available $A_d(C)$. I capture this relational prominence by defining:

$$R_d(C) = \Delta_d(C) A_d(C), \quad (2.14)$$

Although $\Delta_d(C)$ in itself provides a measure of contrastiveness as defined by the distance between Figure and Ground by our definition from Chapter 1 as a measure of dissimilarity, which is fitting to the scale provided in Figure 1.2, this contrastiveness measure in itself is not sufficient to decide for verbal guidance. Had our goal been to only estimate the contrastiveness and locate it on the scale, the $\Delta_d(C)$ alone would have sufficed. However, our goal is to not only estimate contrastiveness but also find contrastive contexts where negations can be used. In this case, the clear alternative component $A_d(C)$ was added to ensure that guidance of explicit contrast (such as negation) remains valid (Singh & Rohlfing, 2025a). For an example, consider a context reflecting simple locative contrast such as, ball is *on* the box, here even though the “beside” is a possible location, our model would not use negation in this case since the context neither provides a “real” alternate nor deviates from ground (low distance). Hence the location now on our contrastiveness scale would jointly be estimated by both the components $\Delta_d(C)$ and $A_d(C)$ as relational prominence $\mathbf{R}$. The decision rule for negation however is a part of next section, where I try to formalize the contrastive guidance based on provided scale. Before getting into the decision rule it is worth noting that we are already at the stage of contrastiveness calculation in the model. The whole model-flow is depicted in Figure 2.9. The model shows the process flow until the contrastive computation takes place (Equation 2.14). After computation, the Guidance module selects the decision ($Neg_d(C)$) for using negation depending on the $R_d(C)$ value, for which the calculation is shown in Equation 2.15.

Decision rule Having laid down our contrastiveness formalizations in Equation (2.14), a simple starting point for contrastive guidance could be to check if the relational prominence $\mathbf{R}$, given by $R_d(C)$, is greater than any given threshold θ_d . In such a case, the model would negate the Ground $G_d(C)$ (Luce, 1959). In such cases, negation is selected whenever relational prominence $R_d(C)$ exceeds a threshold θ_d . In practice, however, such boundaries with hard rules are sharp, as given by Equation (2.15).

$$Neg_d(C) = \begin{cases} 1 & \text{if } R_d(C) \geq \theta_d, \\ 0 & \text{otherwise,} \end{cases} \quad (2.15)$$

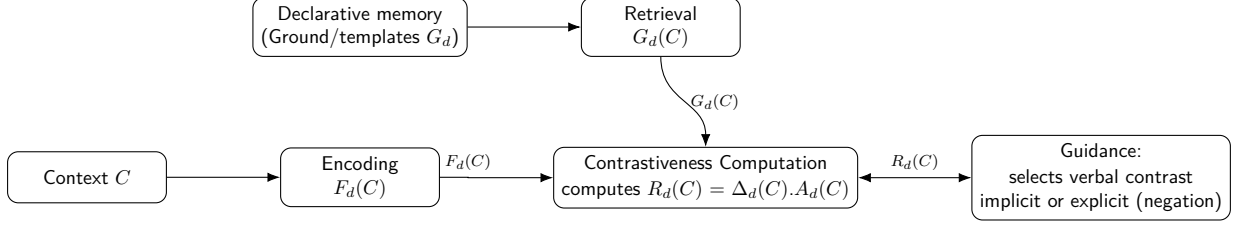


Figure 2.9: Schematic ACT-R style model of contrastive guidance (ContACT-R). The current context C is encoded as dimension-specific chunks $F_d(C)$ (Figure). Declarative memory provides canonical templates G_d (Ground), which are retrieved as $G_d(C)$ for the current context. A comparison stage computes distances $\Delta_d(C)$ and relational prominence values $R_d(C) = \Delta_d(C)A_d(C)$, where $A_d(C)$ captures the availability of alternatives. These values together feed into a guidance policy that in a context chooses between implicit devices and explicit contrast with negation ($Neg_d(C)$).

Hence, to capture graded behavior, I consider a probabilistic version in which the probability of selecting negation increases monotonically with $R_d(C)$, such as captured through a logistic function (Sutton & Barto, 2018):

$$P(\text{Neg}_d(S) = 1) = \sigma(\lambda_0 + \lambda_1 R_d(C)), \quad (2.16)$$

Here, λ_0 and λ_1 are free model parameters that can be estimated from experimental data, and the decision boundary created by the model is shown in Figure 2.10.

Nevertheless, one question that naturally arises at this point is: if the decision to use negation is met, which constituents are to be negated and how might such a construction with respect to Figure and Ground look like? Although we have treated the decision criteria as a hard rule with Negation—either ‘Yes’ or ‘No’—the contrastiveness scale is a continuum taking a continuous value from 0 to 1. Such a bounded decision can be modeled probabilistically using a logistic rule (Luce, 1959). Hence, we see the decision rule $P(\text{Neg}_d(C) = 1)$ as a continuous function of $R_d(C)$ bounded between 0 and 1. With respect to the instruction, this graded decision rule can be visualized as in Figure 2.10. On the x-axis, we see the relational prominence $R_d(C)$, and on the y-axis, $P(\text{Neg}_d(C) = 1)$, where the probability of using negation becomes higher as the relational prominence increases. Crucially, the decision rule also incorporates the different ways of using negation (or no negation) depending upon the value level of contrast, such as low, mild, and high. These cases with examples are elaborated in the following section.

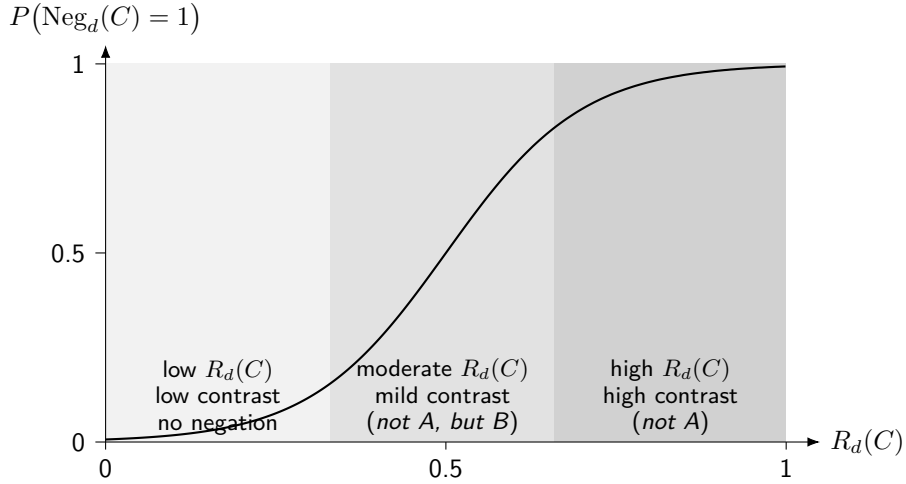


Figure 2.10: Relational prominence $R_d(S)$ as memory-based distance on the contrastiveness scale, and its mapping to guidance choice. At low values, the scene is script-consistent and implicit devices suffice. As $R_d(S)$ increases, alternatives become live and explicit contrast with negation becomes increasingly beneficial, especially for highly contrastive, alternative-rich cases.

2.5 Contrastive Guidance

Contrastive guidance, as I use the term here, is the deliberate use of linguistic cues to *foreground relations* between Figure and Ground so that an observer encodes not just *what* is acting but *how* it stands in contrast to contextually salient alternatives (Singh & Rohlfing, 2025a). In the previous sections, I have argued that contrastiveness is best understood as a graded property of a scene, that is, a matter of how far the realized Figure–Ground configuration is displaced from a canonical or expected template along the contrastiveness continuum. As a starting point, I have also tried to capture and formalize these graded characteristics to adjust how contrastiveness can be assessed and how appropriate verbal guidance can be used accordingly. As described in the previous section, language—and especially explicit devices such as negation—has the potential to either move a scene along this contrastive continuum without changing the underlying visual display by recruiting alternatives and changing which part of the event structure is taken as ground, or can complement contrastive perception itself (Kaup et al., 2007). Building on the contrastive contextualism introduced in Chapter 1, I shall treat guidance as a mapping from a scene’s *relational prominence* to a choice of *implicit* versus *explicit* linguistic contrasts. These choices, as we’ll see, are an outcome of relational prominence (R), capturing the Figure–Ground distance on the contrastiveness scale (Figure 2.10). Here, a low R value corresponds to normative, “script-consistent” configurations, and higher values correspond to non-normative or intentionally opposed configurations. Crucially, the distinction is not either-or but

stays on a graded continuum depending on the available experiential Ground.

The empirical synthesis in Section 2.4 showed that explicit contrast—in particular, negation used in a plausible, contrastive context—can raise this prominence R by (a) making alternatives available and verbally highlighting the ground (Nordmeyer & Frank, 2014), and (b) inhibiting prepotent routine responses caused by the attributes of the Figure (de Vega et al., 2016). In other words, contrastive guidance can be a linguistic “control knob” that adjusts how strongly a given Figure–Ground relation should stand relative to each other, to encode what has happened before in relation to what is happening now. Hence, the goal of contrastive guidance is not simply to negate an already-formed event representation, but to negate in order to allocate attention across the event structure and within dimensions such that the comprehension of relations is optimized. These relations are characterized by low, mild, and high scales, and based on the relational score, we shall provide a way to use negation in such a way that it complements already established representational formats. Hence, for each of these levels (low, moderate, and high), I lay down a linguistic construction, providing a rationale for why it would suit a particular level, with examples.

These examples are importantly inspired by past studies on action performance showing that observers are naturally drawn to the final goal (prepotent figural response) and tend to under-encode the differentiating mid-event segments when left unguided (Lakusta & Landau, 2005; Singh & Rohlfing, 2025a). Introducing contrastive verbal guidance, especially negative sequences for contrastive events, shifts this balance. Seen through the present lens, these findings indicate that explicit verbal contrast increases R for the targeted relation at different levels as provided hereunder:

From relational prominence to guidance choice We denote relational prominence (Figure–Ground distance on the contrastiveness scale) on some specific dimension and context by $R_d(C)$. Empirically and theoretically, the previous sections already provide some ideas on a graded decision rule for the contexts in which contrastive guidance should be deployed. These conditions can be understood by taking examples from each scenario:

Low $R_d(C)$. At the low end of the scale, relational prominence is not explicitly offered by the context but stays there to be extracted; in such cases, the use of negation might not be a favorable choice due to the Figure anchoring perception more than the Ground, and hence the ground seldom comes into attention due to the lack of a clear alternate (Nordmeyer & Frank, 2014). Thus, in other words, the scene roughly conforms to a familiar script and there is no strong pressure to oppose it to an alternative. For example, an event described as, “a ball rolling *into* a box” without any strong alternate on the locative dimension. In this case, negating the already less activated Ground (alternate) does not make sense (such as saying the ball is not rolling *outside*).

Hence, under these conditions, the model should avoid using negation and rather rely on implicit guidance through highlighting the *change in ground*, which comes under implicit forms of negation as discussed in Section 2.3 (see also Table 2.1). In such cases, highlighting the change in Ground through *implicit* linguistic contrasts, such as lexical Path choices (enter vs. move in), resultatives, causatives, aspectual choices, and focus marking, is often sufficient to gently nudge attention toward the relevant relation without overtly invoking alternatives (Talmy, 2000). In this particular example, the linguistic construction profiles a particular boundary or outcome (into the box) but leaves the competing options (if there are any) in the background. Hence, negating in this situation would lead to an “unsolicited” raise in R and could even introduce implausible alternatives that are not supported by the context (Watson, 1965). The Low $R_d(C)$ case is shown in the extreme left part of the decision plot (Figure 2.10).

Mild $R_d(C)$. The mild relational prominence is a case where the Figure and Ground stand for each other without providing an explicit alternate out of many likely alternates. Such cases are generally marked by the availability of the possible alternate(s) such as a possible “Ground”; however, this alternate does not entirely compete for the Figure’s agreement. On a locative dimension, it could be a case where a book is kept in a slot out of multiple shelf slots where the other possible slots are live alternatives, and hence such relations are “almost canonical” placements. In other words, in the mid-range, there is some friction between routine expectations and the current relation, such as the Figure is still at a plausible place, but the configuration is less canonical (e.g., choosing one slot over an adjacent one). Here, the scene already furnishes a natural small set of alternatives, and the task of guidance is to make this set explicit enough so that the observer encodes *which* relation matters (e.g., support vs. containment, first vs. second slot) while preserving the flow of the demonstration. An obvious way to guide under these situations is to negate the prepotent Ground and provide the alternate in turn so that the alternate is also made explicit within the frame of action. This is captured by forms such as “not A (prepotent Ground), but B (alternate)”. Linguistically, these types of guidance acknowledge that an alternative is live and gently steer the observer toward the intended configuration, increasing R just enough for the contrast to become part of the event schema (Horn, 1989).

High $R_d(C)$ (implausible or intentionally contrasted). As stated in the formalization and also understood from examples, the high relational prominence case emerges when two crucial requirements are met: a) a prepotent ground is already there, i.e., a high implausibility where the Ground originating from experiences becomes instantaneously active—this case can be best understood in relation to the situative contrast, such as when the *bin is placed on the crossing*; and b) there is a true alternate (perhaps only one prepotent) that is competing with the current figure’s existence (Garfinkel, 1967). As a result of this sustained conflict, the Fig-

ure–Ground relation becomes more prominent than merely the figure’s attributes. At this high end, alternatives are prominent and the existing Figure–Ground relation is either implausible relative to the experiential script or deliberately set up as a contrast (e.g., a bin placed on a crosswalk). In such contexts, the visual or situative layout already provides a strong ground for comparison. Since the situative dimension is mostly descriptive rather than instructive, we might usually observe this ground prominence at the verbal level as, “the bin is not on the *roadside* (canonical Ground)”. For explicit guidance, we must rather look at the eventive or locative cases which are more instructive. In an event, if the alternate is already pushing itself forward, negating only the past (for eventive) Ground is enough to bring together current and past events into attention. Studies show that when negations are used on the *later* subevent for eventive contrasts, they reference the current subevent and prior alternative simultaneously, thereby raising R at the precise boundary that differentiates the actions (Singh & Rohlfing, 2025a). For example, experiments as reported in Section 2.4 on assertive–negative guidance demonstrate that for contrastive sequences with a clear alternative, negating the later subevent improved recall for the differentiating subevents. In this sense, guidance is linguistically contrastive, such that it asks the observer to see the past event (the one that just occurred) *through* the lens of the current event, bringing them together in a single interpretative framework.

The levels of contrast at linguistic scales together bring guiding strategies that depend upon how the different dimensions come forward to bring constituents in relation to each other. Contrastive guidance hence extends the earlier claim that language can shift perception on the contrastiveness scale and also be used to highlight contrasts when the context provides such opportunities. The present section specifies how these devices can be deployed in a principled way. By aligning the choice of implicit versus explicit contrast with the current level of relational prominence R , and tuning it according to the context, verbal guidance can systematically redistribute attention across the event structures, supporting encoding of Figure–Ground relations without altering the existing configuration itself. The subsequent section will take up this strategy in more applied settings by outlining some principles for contrastive guidance based on context and alternatives. Whereas the entire question of context is a dedicated discussion of the next chapter, in the subsequent section I will extend some light on the question of alternates and their use for contrastive contexts.

2.5.1 Design principles for contrastive guidance

The empirical pattern from the action studies suggests a set of simple, transferable design principles for using negation as a contrastive tool in instruction and demonstration. These principles are deliberately phrased in task-agnostic terms related to the contrastive dimensions. However, each of these principles instantiates the same

basic idea: to raise contrastiveness on the Figure–Ground scale by negating an appropriate ground, inhibiting a prepotent routine, and providing the guidance within an appropriate temporal slot (synchrony).

Negate where an alternate is live. Negation is most effective when it does not “float in the void” but opposes the current subevent to a just-established alternative (Wason, 1965). Crucially, the availability of a single, clear alternative is necessary for the negation to implicitly bring the two events within the same frame of reference (Orenes et al., 2014). Such is the case of Up vs. Down, where only two possible locations are possible during the entire event. For such locative and eventive contrasts, guidance is naturally achieved by placing negation for the later subevent to establish a rich context (such as, Up—*not*-Up), so that a concrete prior event can serve as ground. In practical terms, this means first establishing a clear candidate pattern, then using negation to cancel or deviate from it. This late-position negation was precisely the configuration in which we observed the strongest boost in mid-event encoding for contrastive actions (Singh & Rohlfing, 2025a).

Pair negation with assertion. Consider the example from our experiment where the assertive and negative guidance come together. Such a construction is necessary to establish a positive counterpart before negation can act on it (Kaup et al., 2007). Hence, pure negation without a credible counterpart risks remaining an underspecification. Aligning with this, if we had used the negated guidance at a later subevent—e.g., *No instruction* → *not Up*—pragmatically it would still be a feasible framing since the previous path was opposite to the current one, and implicitly the visual direction alone would establish the “Upward” event. However, the effect might not have been as impactful as one would get in a verbal framing like “*Now Up* → *Not Up*”. Hence, establishing a positive counterpart at the multimodal level (verbal-visual) is important to make the contrast more explicit. From the contrast-as-context perspective developed in Chapter 1, a well-formed use of negation should make the relevant alternative explicit or at least strongly implied. Similarly, other constructions, such as the form “*Not A, (but) B,*” do exactly this: they both inhibit a prepotent or misguided routine (A) and supply a workable alternative (B) as new ground. In our terms, this strengthens alternative availability and avoids leaving the learner with a gap. Even when B is visually present, articulating it verbally alongside the negation can help align linguistic and visual processing, a reason ascribed to multimodal facilitation (Bahrack & Lickliter, 2000).

Matching the locus of guidance to the task. When we talk about multimodality, a crucial aspect that comes into the picture is the idea of multimodal synchrony, which was shown to modulate perception, as discussed in Section 2.5 (Singh & Rohlfing, 2025a). Hence, presenting guidance is not sufficient; it should be applied a) at the level that is functionally relevant for the contrast in question, and b) with specific timing. Functionally relevant constituents in an action key distinction concern

the elements of action, such as Path vs. Manner (Talmy, 2000). When the *where* of an object is under focus, negating Path is the natural choice, so that it redirects trajectories and supports goal contrasts on the locative or eventive dimension. When the distinction concerns *how* an action should be carried out (e.g., hit vs. press, push vs. twist), negating Manner is more appropriate, such that it suppresses an ill-fitting routine and highlights the fine-grained realization that matters for functional success. Both cases are instantiations of the same general mechanism, but they differ in which Figure–Ground relation is being pushed rightward on the contrastiveness scale. With respect to multimodal synchrony, the timing of the presentation can make certain elements more salient than others, and hence it should be used at an appropriate synchrony as suggested in Figure 2.7 (Gogate & Bahrick, 1998).

The known caveats Taking a modeling perspective has the usual double edge. On the one hand, a model forces the theory into a more concrete and applied format; on the other hand, it inevitably builds in rigid assumptions that may not generalize to all cases one cares about. One such point of caution here concerns how well the current formalization can accommodate the scenarios actually created by different users. In terms of contrastive perception, we should ask: how well does the present account capture *individual* sensitivity to contrast? As already noted in Chapter 1, the extremes of the contrastiveness scale are likely to be relatively stable across a population: highly normative or maximally implausible situations tend to be perceived in broadly similar ways. However, in the middle range of the scale—where many of the interesting everyday cases live—individual variability is not a nuisance but a crucial part of the picture, especially if the goal is to tailor verbal guidance to specific learners or users. The current model treats relational prominence $R_d(C)$ and the corresponding decision rule in a relatively uniform way; in practice, thresholds and alternative availability are likely to differ across individuals and contexts (Lovett et al., 2000). This limitation is acknowledged and should be addressed in future work by allowing user-specific parameters or experience-dependent tuning rather than a single, population-wide policy.

A second limitation of the current formalization is that it is effectively *atemporal*. The model, as laid out, evaluates contrastiveness and selects implicit vs. explicit guidance, but it does not incorporate instruction timing as a first-class variable—even though multimodal synchrony was shown to affect attention to action constituents (cf. Section 2.3). As we have seen, instruction–action timing changes what the instruction *does* for the learner. When guidance precedes the subevent (*instruction-before-action*), it functions as an ostensive marker, announcing that “something relevant is about to happen” and shifting attention away from the goal toward the upcoming mid-event (Singh & Rohlfing, 2025a). By contrast, when guidance overlaps with the unfolding action (*instruction-along-action*), verbal and visual information are packaged into a more unitary, goal-oriented percept (Bahrick & Lickliter, 2000). In this overlapping

regime, negation still carves out contrast, but its role becomes more selective: overlapping simple assertions best support noncontrastive sequences, whereas overlapping assertive–negative patterns are particularly suited for contrastive ones (e.g., Over–Under), where an A–not A format can keep the alternative in play even as perception is drawn toward the goal (Singh & Rohlfing, 2025a).

Instructional designers can therefore treat synchrony as a parameter: use sequential cues to generally spotlight subevents, and overlapping cues when the aim is to embed an A–not A opposition into an already ongoing, goal-directed routine. At present, this temporal dimension is only reflected in the informal “policy” for contrastive guidance, not in the model itself. A more complete future version should integrate synchrony into the formal account, so that the choice of when to speak becomes as systematically modelled as the choice of what to say.

2.5.2 Discussion

As the current chapter has shown, language does not merely act as labels for already individuated objects and events; rather, it can structure relational perception and act as a contrastive mechanism (Talmy, 2000). This is achieved by meddling, spotlighting, and augmenting Figure and Ground, making the relation between the two prominent while leaving others in the background (Wolff & Holmes, 2011). In this sense, contrastiveness is not a rare property such as experienced in explicitly contrastive sentences, but a graded feature of everyday event apprehension that depends on how far the current Figure–Ground configuration deviates from a canonical situation. Language, through its lexical choices, syntactic packaging, prosody, and especially through explicit contrast devices such as negation, can then be used to steer this relational prominence in systematic ways (Talmy, 2000).

The empirical study by Singh and Rohlfing (2025a) supported this idea by treating event perception as a dynamic process where linguistic guidance can meddle with ongoing conceptualization. Motion events were a natural starting point because they already had a Figure, a Path, and a Goal forming the primary test-bed for contrastiveness (Talmy, 1985b). In Singh and Rohlfing (2025a), it was shown that contrastive verbal instructions—where negation such as “*Now Over – Not Over*” made the present subevent to be interpreted in light of what had just happened—foregrounded the relation between the two paths rather than the isolated path, leading to better event memory. Furthermore, the eye-tracking and recall data converged on the idea that goal bias, which is a very robust finding in event perception, can be attenuated when language explicitly builds a shared representational ground that includes both past and present subevents (Lakusta & Landau, 2005). In this way, negation helped to attenuate attention from the event-end goal back to the mid-event structure.

Alongside these motion event studies, the work on manner learning in Human–Robot

Interaction extended the same contrastive framework to the functional dimension (Groß et al., 2023). Everyday objects often come with highly routinized affordances, such as a button that is every time meant to be pressed (Gibson, 1979; Norman, 1988). In the touchscreen paradigm, such prepotent manners functioned as an experiential Ground against which an alternative manner could be introduced. Formulations like “*do not press, but turn*” explicitly inhibited the routinized response and at the same time highlighted a less frequent but functionally appropriate alternative. The interesting point here is that the benefit of negation does not lie simply in negating the incorrect action, but in binding the wrong and the right actions into a single contrastive unit so that the learner comes to understand their relation and not just the target response. In other words, negation did not operate as a purely truth-functional operator, but rather as a scaffolding device that increased relational prominence on the manner dimension by bringing an alternative within a functional dimension (Horn, 1989). Although negation is cognitively more demanding than simple assertions, when the context is contrastive it becomes a powerful device to generate rich contexts, where the present action is anchored in light of nearby alternatives (Nordmeyer & Frank, 2014; Wason, 1965). The studies on motion events and on action manners together suggest that negation can simultaneously fulfill two roles: it brings alternatives into attention at the representational level, and it recruits domain-general inhibitory mechanisms to suppress the default response at the sensorimotor level (de Vega et al., 2016). When these two roles integrate, explicit contrast becomes supportive for scaffolding.

A further step in this chapter was to make the notion of relational prominence more concrete by developing an ACT-R model of contrastiveness (ContACT-R). Here the aim was not to build a complete architecture of language and perception, but to show that the informal notion of Figure–Ground distance can be captured in a way that is compatible with existing classical and interpretable cognitive modeling frameworks, such as ACT-R (J. R. Anderson et al., 2004). Although at the initial stage, in the current model, a given context is encoded as a dimension-specific Figure $F_d(C)$, which is then compared with a dimension-specific Ground $G_d(C)$ retrieved from declarative memory. The similarity between these representations, combined with a simple estimate of alternative availability, yields a value for relational prominence on that dimension $R_d(C)$. This value can then inform a simple guidance that decides whether only implicit contrastive devices are needed or whether explicit contrast, up to and including negation, is warranted. In this sense, the model acts as a bridge between the phenomenological descriptions of contrastiveness from Chapter 1 and the experimental findings in this chapter.

Being aware of the limitations of the current model—which purposefully abstracts away from questions such as how canonical Grounds are learned in the first place, how multiple candidate Grounds compete, and how the figure is dynamically assigned when several entities are in motion—a need to bring learning histories, developmental trajec-

tories, and interactive scaffolding was highlighted. This also prepared the foundation for later chapters where contrastive guidance will be embedded into a more fully developed into an interactive (InterACT-R) framework. Another limitation is that the chapter has mostly treated context as something that can be reconstructed from an individual observer’s memory and present perceptual input. Although I will argue that context is both given and developed, the empirical work here has remained closer to the static view. For instance, the Ground was either previously experienced in the same experimental session or provided by a highly conventionalized script that most participants were assumed to share. What is still largely missing is the kind of context that emerges between agents as they coordinate over time and explicitly work to align their perspectives (H. H. Clark, 1996).

These limitations call for a contrastive guidance framework that extends the domain of interaction and scaffolding. If language can serve as a mechanism for structuring how events are perceived in a static context, then in interaction it can be used to structure how a joint activity is negotiated and maintained (Tomasello, 2008). In such a setting, Grounds will not just be prior subevents or canonical templates stored in memory; rather, they shall also be built and revised through the turns that participants take. Relational prominence will then have to be understood not only as a property of an inner representation but as something that is made manifest in how partners orient to each other’s actions and in how they monitor and correct each other. The next chapter therefore shifts the perspective from isolated scenes to interactive settings, in particular to Human–Robot Interaction where a robot not only demonstrates actions but also needs to respond to a human partner in real-time. The methodological proposal that I briefly foreshadowed already, which keeps the robustness of controlled experiments while preserving interactivity, will be the place where the contrastive ideas from this chapter are put to a new test. The plan is to move from ContACT-R, which models contrastive guidance in relatively static scenes, toward an InterACT-R model in which grounds are dynamic and co-constructed by the learner (or explainee) and the explainer (Rohlfing et al., 2021). In that extended framework, negation and other contrastive devices will be treated not only as tools for action clarification but as instruments for scaffolding, monitoring, and co-constructing a partner over time.

Chapter 3

Context

3.1 Why Context?

One way to appreciate how important context is for situational understanding is to think about how utterances should be properly contextualized when the recipient receives them. Consider, for instance, the simple utterance, “*Your coffee, sir!*”. As soon as someone hears it, it automatically presupposes a number of silent conditions to be met before it is felicitously interpreted. Roughly, we assume that (a) The speaker is holding a coffee with an intention to serve it; (b) The addressee has in some sense ordered or is entitled to that coffee and is therefore the recipient; and (c) The action proceeds from server to receiver, i.e., a server–customer relation, perhaps involving tables and a restaurant- or café-like environment (Searle, 1969).

Conversely, if the utterance itself is produced under precisely these conditions as described, there might not be a need to verbalize and elaborate the conditions under which it is taking place. In such cases, there is no need to support the utterance through multiple sentences, and the task of understanding is easier, where a mere single utterance becomes the entire description of the situation. This is perhaps one of the central functions of context – it allows us to avoid all sorts of descriptive complexities and still enables us to convey a rich amount of information through minimal elaboration, such as shown in Figure 3.1. To elucidate this idea with a simple illustration, when context surrounding an event is rich, the event itself can be described in a compact way (Figure 3.1, lower). In contrast, when there is little or no context, the same event would warrant a much longer description (Figure 3.1, upper). Context, in this sense, is a very rich repertoire of things that are already in place, evolving with and supporting the event in question (here: the act of serving coffee). It is important that the speaker and recipient align on a “common ground” before the potential of context-making utterances (compact yet meaningful) can really be exploited (H. H. Clark, 1996).

This leads to a core question that is not only *what context is*, but also *what context does* (Dourish, 2004). From a functional viewpoint, context may not be something that just stands in a situation, a static backdrop of facts, but rather something that is working to make utterances more meaningful by preventing over-complexity. Once

a common ground is established, context allows much of what is relevant to remain implicit. Seen from another angle, context is both brought forward in an interaction to support efficient conversation and, at the same time, is something that utterances require and can *refer* to in order to be interpreted at all (Dourish, 2004). In this sense, actions (including utterances) are “situated” within a context which, in principle, is overwhelmingly rich (not everything that is part of the context can be mentioned) and does not need to be accompanied by explicit verbalization, even though the utterance would not fully make sense without it (Suchman, 1987). Something seems to be working very smoothly in this case, such that language is able, when needed, to make selective use of available context. In the coffee case, an utterance such as, “Your coffee, sir!” becomes entirely meaningful by virtue of being projected back onto the surrounding situation. The interpretation is rich because the visual and social scene act as context for the utterance (Tanenhaus et al., 1995).

One way to push this idea further is to ask: if visual events can act as context for language and action, can language itself sometimes act as a “ground” for subsequent utterances, establishing a context *within* an unfolding event or discourse? In other words, can utterances alone play the same contextual role for later ones that scenes and scripts play for the first? At the most basic level, context should be able to act to make us utter less and say more, and here lies much of its power. It goes without saying that the tremendous amount of information available or made available to us as a part of interaction in a given situation is not all explicitly encoded in the utterance. Conversely, only a few facts are “selected” and become relevant for the situation (Rohlfing et al., 2026). This is the case where the recipient already has a “selected context” (or Ground, as we’ve been calling it) within whose presumptions the utterance is woven by the speaker. But what about the cases where the speaker wants to *selectively* make such a Ground salient and relevant for the utterance?

Consider a slight variation on the coffee example. Suppose the recipient comes to this particular café every day for a tea, but today tea is not available. There is a common ground between the speaker and recipient that, for this person, tea is normally the relevant drink in this situation. Without this ground, the interaction might proceed along the full routine of ordering: the server asking what the recipient wants, listing options, confirming the order, and so on. The most likely or efficient way of handling the current situation would be to list the available drinks and let the recipient choose. However, one can imagine an even more compact construction such as, “*Today, no tea, sir, but coffee.*”, which, importantly, makes the ground explicitly part of the context through the utterance itself.

The crucial difference between the two utterances lies in what is made explicit. Whereas the first utterance can be fully successful and informative, it simply lists the available options and leaves the choice to the recipient. The second utterance, however, does something extra by explicitly bringing more background and shared knowledge into

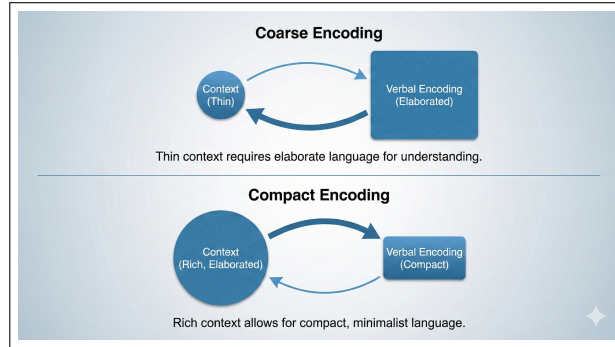


Figure 3.1: Schematic relation between context and verbal encoding. In the upper panel, a relatively sparse context (small circle) supports only a coarse verbal encoding (larger, more elaborated box): much of what is needed for understanding must be spelled out in language. In the lower panel, a rich context (large circle) allows for a compact utterance (smaller box), since many presupposed conditions are already provided by the situation and common ground.

play through *negation*. It does not just say “coffee is available”; it says “no tea”, thereby activating and simultaneously inhibiting what the recipient would normally expect (Kaup et al., 2007).

This is precisely what linguistic negation allows us to realize, something that cannot be achieved by simple recursive assertions. Negation is known to recruit a rich contextual repertoire of alternatives into consideration so that utterances become more informative without requiring that all past information be explicitly verbalized (Horn, 1989; Kaup & Dudschig, 2020a). By saying “no tea” rather than simply asserting “we have coffee”, the speaker acknowledges the shared history of tea orders, marks today as an exception, and positions coffee as the alternative that now matters. Negation, thus, functions as a linguistic constituent that connects the present move to the ground furnished by past interaction.

In this chapter, I will primarily focus on such a functional approach for understanding context. Although the following sections will take up classical questions such as “What is context?”, “What is its nature?”, and “How has it been treated in the literature?”, I also aim to use a substantial part of the chapter to extend the notion of *contrastive contextualism*, which, at one level, can be explained through linguistic negation.

First, I will introduce different notions through which the question of context can be approached. Then, I will particularly motivate two perspectives on how context can be defined, depending on whether we treat it as something available or as something emergent (Dourish, 2004). Finally, I will discuss how context can be seen as a by-product of interactive processes, where it is continually shaped and reshaped by what participants do during interaction. Consequently, each of these sections will be supplemented by empirical results (now published) that form part of the thesis, and I will bring these results in at points where they naturally connect to the theoretical discussion.

3.2 Approaches to studying context

In the literature, the notion of context is usually treated from two broad perspectives (Dourish, 2004). On the one hand, context is treated as a representational problem. Here one generally asks: how is context internally represented? Which cognitive processes operate on those representations? And, how do these processes give rise to behaviors observed on the linguistic and nonlinguistic levels? On the other hand, context is seen as a fundamentally social and interactional phenomenon: here the emphasis is not on what context *is* in advance, but on how something *becomes* contextually relevant for a particular situation. In this sense, context is emergent and an outcome of social interaction (Schegloff, 1992). Crucially, this distinction is not merely terminological; rather, it traces back to different scientific traditions in the social sciences. For the present discussion, I take two relevant notions into account: the positivist and phenomenological approaches (Dourish, 2004).

As far as positivist theories are concerned, they are typically derived through rational and empirical methods, where complex observable phenomena are reduced to measurable variables which are often formalized in mathematical descriptions. These observational units usually form the basis for representations that are studied empirically. Thus, in the positivist tradition, the central goal is to gain objective knowledge and context-independent descriptions of social or cognitive phenomena by abstracting away from the particular situation in which they occur (Winograd & Flores, 1986). On a broader note, the aim is often to derive statistical or computational models of the phenomenon in question, so that the processes and representations behind it can be objectively investigated. For investigating context, this means asking questions such as: What are the representational structures that constitute “the context” for an action or utterance? How are these structures stored, retrieved, and updated when cognitive processes operate on them? Akin to this approach, a representational account of context occurs with the model in the previous chapter, where the contrastiveness scale, the relational prominence $R_d(C)$ in the ContACT-R model (Section 2.5) all presupposed that context can be at least partly captured as structured information in memory that is available to be consulted and manipulated through cognitive operations.

In contrast to the positivist stance, phenomenological approaches seek a qualitative description of the phenomena in question. They resist reducing social facts to objective quantities and instead treat them as fundamentally subjective, i.e., having no independent reality beyond what is perceived, negotiated, and enacted by participants (Dourish, 2001). In this view, social facts are emergent properties of interaction and are continually re-shaped in processes of interpretation. Phenomenologically, we recognize the world as we perceive and live in it, pushing us against the idea of a fully objective and stable external world that is merely “represented” inside the head. For contextual analysis, phenomenological and interactional theories (e.g., conversation analysis,

ethnomethodology) would argue that what we call “context” is not simply an abstract representation attached to a situation, but is actively *constituted* in and through interaction (Goodwin, 2000). Thus, context is not something that is simply “there” before the interaction starts; rather, it is partly produced by the very actions that are supposed to be situated in it (Dourish, 2004).

Seen from this perspective, the abstract representations that a positivist account postulates are themselves subject to interaction; such as which aspects of a setting become relevant, which background expectations are invoked, and which histories are treated as live depending on how participants orient to each other and to the task (Rohlfing et al., 2003). Following this, what counts as context for an utterance in our example “Your coffee, sir!” is not just a pre-given list of features (a table, a cup, a waiter), but also the ongoing activity (serving, receiving), the shared script (restaurant service), and the way in which the partners momentarily align to that script. From a phenomenological standpoint, context is thus dynamic and indexical, where its development is an intrinsic part of the unfolding interaction rather than fully specified in advance.

For the present thesis, both perspectives are methodologically necessary. The notion of *contrastive contextualism* developed in Chapter 1 and its formalization in Chapter 2 treat context primarily as representational, where experiential grounds, canonical templates, and relational prominence are assumed to be available as retrievable structures. At the same time, the later chapters on scaffolding treat context as an interactional achievement that is actively shaped through communicative moves (e.g., negation, repair, and monitoring). To make this bridge more explicit and to capture the form of context needed for the later analyses, I adopt the systematization of context in social XAI (sXAI) recently proposed by Rohlfing et al. (2026). In this framework, context shifts from a static conceptualization, as a set of social and material facts, to a *co-constructed* context. Unlike a purely representational account, in which social and material facts are (pre-)selected to characterize a situation, a co-constructed context is defined as a dynamically evolving set of relevant facts and commitments that emerges through the unfolding interaction and is continuously shaped by the partners (Rohlfing et al., 2026). In this sense, the goal is not to privilege one *perspective* over the other, but to connect (a) what is available to cognition at a given moment with, (b) how relevant constituents within a context is adjusted and co-constructed during interaction.

3.2.1 Context is what is available

This section defines what *available context* means in this thesis: a structured set of representations that can serve as *Ground* for interpreting a current *Figure*. On a representational view, context is treated as information that is already there and can be drawn upon to characterize a situation (Dourish, 2004; Minsky, 1975). In the terms introduced in Chapter 2, this corresponds to the idea that a current configuration

(Figure) is interpreted against retrievable templates or patterns (Grounds), and that contrastiveness depends on how the Figure matches or departs from these Grounds.

For example, the restaurant situation depicted in the beginning of this chapter also utilized this mapping where an utterance like “Your coffee, sir!” can be understood as felicitous because participants can rely on a service script and role structure (server/recipient), together with a locally ongoing episode (an order having been placed). From the modeling perspective, in the ContACT-R account, such information is treated as Ground content that is already active or easily retrievable, so that the utterance can be interpreted against it.

Although hard coding such symbolic representations reduces the complexity, defining context as “what is available” raises two immediate limitations. First, availability is in principle unbounded: almost anything in the vicinity of an event could be counted as context, yet only a small subset ever matters for interpretation, yielding a selection problem (Minsky, 1975). In ContACT-R model, this was handled by cue-based retrieval, i.e., given a current context C , the system searches for the most appropriate Ground $G_d(C)$ on *particular dimension*, so that only content matching the cue becomes relevant for interpreting the Figure. Second, the “available” view risks treating context as static, as if it could be listed independently of ongoing activity. Being aware of these limitations, in the next section, the present notion is made more feasible by specifying the nature of *what* can be available as Ground, structured along the relational dimensions as introduced in Chapter 1 (situative, functional, locative, eventive, social). In subsequent sections, I then turn to a more holistic notion of context is not only retrieved but also developed through ongoing activity.

3.2.2 Content based

Often while addressing the question of *static vs. dynamic* contexts as a difference between contexts that is pre-specified *available* Grounds and contexts that are *reshaped* as activity unfolds (Dourish, 2004; Suchman, 1987; Goodwin, 2000), we are likely to overlook the question of what might be the nature and content of context *per se*. With an initial assumption, one could argue that these representations are in themselves the very entities that we perceive, and hence these representations are isomorphic. Although in a recent review there has been a suggestion to reduce this ambiguity along the nature of representation, such as modal vs. amodal (for a review see, Kaup et al., 2024), the question of representational *content* remain a potential area to explore further.

A promising approach is to characterize representations in terms of relational *dimensions*, particularly when the goal is to reduce the scope and complexity of the problem. This leads to the next question: what kind of *content* does the “available context” comprise? As I have argued in Chapter 1, context is not an amorphous cloud of background information but is structured along specific conceptual dimensions. Crucially, we have

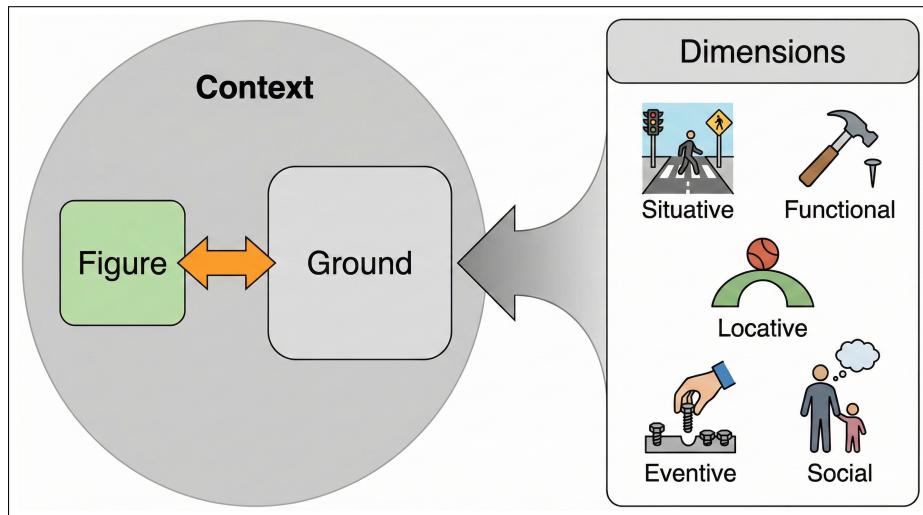


Figure 3.2: Schematic of content-based context. The diagram illustrates how context encloses the Figure–Ground relation. Crucially, the content of context is not arbitrary but structured along specific dimensions (right panel: situative, functional, locative, eventive, and social), which determine which Ground is available for interpreting the Figure.

already utilized these dimensions to characterize contrasts in the previous chapter: situative, functional, locative, eventive, and social. Thus, each dimension contributes to a particular schema along which the type of content can be described. Under this assumption, the Ground acts as a restricted scope for interpreting a current Figure, and each type of content has its own profile of accessibility as shown in Figure 3.2. In order to characterize these dimensions, I will now go through each of them to visualize them concretely.

The **Situative dimension**, at the broadest level, captures the coarse “where and what” relations of a situation; for instance, in situations like a street with bins and curbs, a café with tables and cups, or a lab with screens and response boxes. In representational terms, this is script- and scene-knowledge about typical configurations (where bins belong, what and where objects are canonically placed in a restaurant, or how a participant sits in front of a monitor) (Schank & Abelson, 1977). When we say “The bin is not by the curb”, we are drawing on situative content that encodes normative placements and their roles in everyday routines. This content is often highly accessible and built up over many encounters and experiences, therefore becoming readily available as Ground within the context.

Unlike “contents” within the situative dimension that characterize situations as a whole, **Functional contents** are more fine-grained and concern *how* objects are typically used (Gibson, 1979). They typically answer which actions go with which tool and which manners of acting are licensed or expected (Gibson, 1979). For instance, drawing on

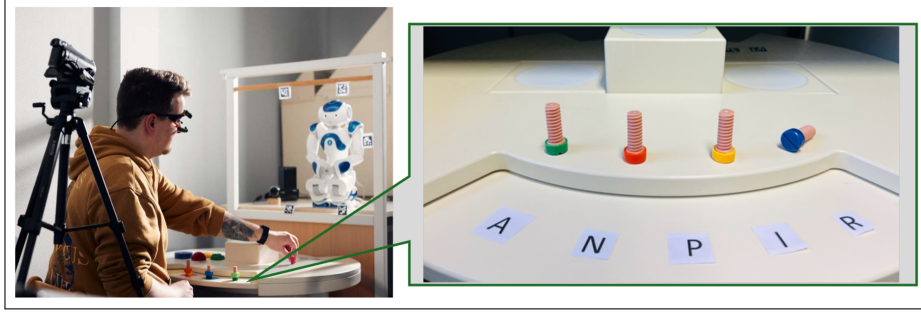


Figure 3.3: Experimental setup for investigating the role of eventive context in Human–Robot Interaction (adopted from (Singh et al., 2025)). The left panel shows a participant interacting with a robot while gaze is recorded. The right panel depicts the workspace in which a sequence of object placements (screws) establishes an action history, or Ground (e.g., a repetitive vertical manner), against which the target action is performed and interpreted.

findings from the study by Groß et al. (2023), instructions such as “Don’t press—but turn!” presupposed a stored pairing between a knob-like object and a default action (pressing), even if that default was in fact misleading and not the correct action. Thus, in this case, the context holds information about *affordances* and action–goal schemas driving which routines are practiced, which alternatives exist, and how they relate to the object (Norman, 1988). In our ContACT-R model, the contents are encoded in the form of manner relations, whose activation levels will influence which manner comes to mind first after the object is encountered.

Locative content concerns the spatial layout of possible placements and paths. Taking example from motion event, for the ball trajectory, this can mean knowing the possible locations a ball can travel: Up or Down, Over or Under, or that it can end up in the box or on the rim. In more everyday terms, it includes possibilities such as “on the chair, under the table, beside the box,” and so on. Representationally, this content often comes in the form of structured spatial templates or frames in which different plausible alternatives (inside, support, placements) are available (Landau & Jackendoff, 1993). Taking an example from Singh and Rohlfing (2025a), when we say “Not Up” or “Not Over, but Under”, we are exploiting structured content at locative dimension with alternates, and the negation works because an alternative location is already encoded as part of the same ground.

Whereas locative content brings together spatial alternates into a frame, **Eventive content** ties these pieces together over time. It captures histories of how an event has unfolded so far and which subevent is most likely to occur given the ground. The content could then have an order, a snapshot of expected continuation, and a goal representation (Zacks & Tversky, 2001). In Singh et al. (2025), such contents are encoded as action sequences of manners forming a context. A similar categorization

could be imagined as rich vs. simple contexts (see next section). Here, the first few subevents lay down an expectation that can be continued, repeated, or violated. From a representational perspective, context here can also include a partial event encoding, such as “locative event Up has just happened”, “vertical manner has been performed thrice already”. Negation can then recruit this context history as Ground and interpret the current subevent as a part of previous actions (Singh & Rohlfing, 2025a).

Finally, **Social content** concerns participants, roles, and scripts. In the restaurant case, context includes who is the server, who is the recipient, where they are in the script (menus, order, eat, pay), and importantly which moves are licensed at which point. Any utterance would then depend on this content, and it will only make sense when there is a shared understanding and social script (Searle, 1969). Representationally, this may involve role labels, turn-taking expectations, and normative sequences of actions. From a contrastive point of view, social content provides the ground on which misalignments can be repaired (Schegloff et al., 1977).

A useful way to summarize these strands is to say that content is both *typed* and *selective*. Typed in a way that different dimensions contribute different kinds of information (scene layouts, affordances, spatial roles, event histories, scripts). Selective, in that not all stored content is equally accessible at a given moment. In the modelling work, this selectivity was implemented via retrieval cues (J. R. Anderson et al., 2004). Meaning, given a context C , only those chunks whose content matches the current cue (e.g., “restaurant service”, “possible path”, “manner relation”) are likely to be activated and serve as Ground. From the perspective of contrastive guidance, this matters because negation can only raise contrastiveness $R_d(C)$ along dimensions for which suitable content is already in play. For example, verbal guidance like “Not Up” would fail if no vertical alternative has been established. Similarly, guidance on the functional dimension like “Don’t press – but turn” would fail if the learner has no representation of “pressing” as a plausible action in the first place.

The content-based view of context thus gives us a more systematic representation of what “available context” actually consists of. Without this conception, these were taken as some arbitrary features around an event, but the contextual analysis allows us to see it as a layered structure of situative, functional, locative, eventive, and social information. In the following sections, I will argue that although dimensions provide us a concrete way to characterize the content of the context, at a more abstract level we can also define their nature based on the informational content regardless of the dimension under which it falls. One such idea has been used in the study by Singh et al. (2025) to test the effect of two different types of context on language and visual processes.

Two types of available context From the content-based notion of context, it can be argued that although the *kind* of content in the context will differ depending on the

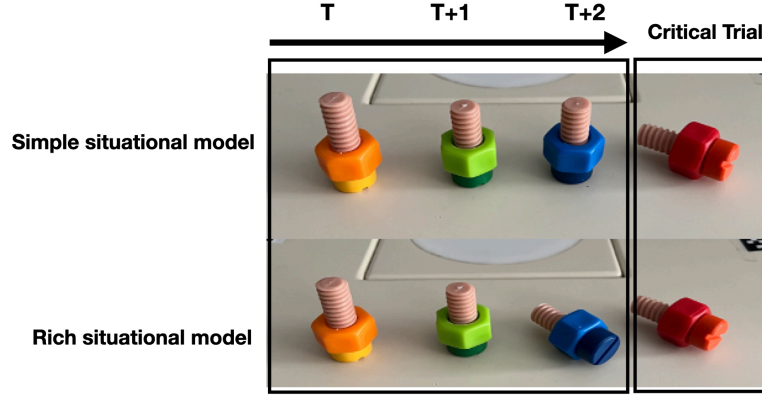


Figure 3.4: Visualization of the two types of eventive contexts used in the study (Singh et al., 2025). The top row depicts the Simple situational model, where the action history is uniform (all screws placed vertically), establishing a single, unambiguous ground. The bottom row depicts Rich situational model, where the history includes mixed manners (vertical and horizontal), thereby keeping multiple relevant alternatives available in the context before the critical trial.

dimension in which it falls or the events that take place (situative, functional, eventive, and so on), the *amount* of information provides an additional way of characterizing contexts (Shannon, 1948). Put simply, some contexts offer a dense, informative ground, while others only sketch the minimum needed for an action to go through. For instance, when all the possible entities relevant for an action are already in place, the situation furnishes a rich repertoire of ground. In the coffee example from the beginning of the chapter, if all the conditions are met—such as the restaurant setting, server–recipient roles, a placed order, and a shared script—then the utterance, “Your coffee, sir!” is embedded in a rich context and can be extremely compact (Levinson, 1983). By contrast, if only some of these ingredients are present (say, there is a cup but no recognizable service script or addressee role), the same utterance would be under-informative and require further elaboration. In this sense, rich contexts are those in which the relevant alternatives are explicitly or implicitly represented. Conversely, simple contexts encapsulate only a subset of the entities or relations that could, in principle, matter for the action. This distinction becomes useful when we want to talk about context informativeness, i.e., not just what type of content a context contains, but how many relevant options it keeps live as a part of potential Ground.

A simple example from the eventive dimension comes from the experiment in which the informativeness distinction was first utilized (Singh et al., 2025). Recall that eventive contrast depends on what has already happened and how this history shapes expectations for the next action, such as placing screws sequentially with the same or different manner (vertical or horizontal) so that the ongoing pattern can serve as Ground for

the upcoming move. In the study reported in Singh et al. (2025), two kinds of eventive contexts were implemented by varying action histories for a sequence of objects.

In the **simple situational model**, all prior actions before the critical action revealed the same manner (e.g., all vertical or all horizontal). This produced a clear, but one-sided ground where the participant only experienced one way of placing the screws, and thus only one manner was strongly available as a candidate continuation. In the **rich situational model**, by contrast, the preceding actions were mixed such that both manners (vertical and horizontal) were represented before the final action was performed. Here the ground encoded not just a representation such as “this is how we have done it so far” but also “these are the possible ways in which it could be done”, making it richer by revealing the full space of manners before the final action takes place (Figure 3.4). Using an eye-tracker, we measured participants’ gaze behavior online as they faced a missing-information trial in the instruction (missing manner information) and subsequently performed the final action after clarification.¹

As we predicted in the study, participants’ gaze behavior showed that the rich contexts (the situations depicting both possible actions) drew more attention than the simple contexts (Figure 3.5). In other words, when all the contextual content pertinent to the current task is explicitly present in the event history, participants are more likely to refer to that history and treat it as a resource for resolving the current task. This fits well with the broader idea that rich contexts are not only dense in terms of content, but they also function as a ground for the current task. In terms of contrastive guidance, such contexts also create plausible grounds for contrasts because they keep multiple relevant options available for being opposed, negated, or completed. As we have pointed out that a distinction between a rich and simple context as such provides a way to quantify the amount of information being held on the eventive dimension, the distinction can also become useful for other dimensions where informativeness is under investigation. Moreover, we established that a rich history allows us to use it as an opportunity to refer back to the contents when we need it for the task, as opposed to simple contexts which only partly provide this support, which we term as a notion of “opportunity” and discuss in the following section.

Context providing opportunity So far, context has been treated largely in terms of what it *contains*: situative, functional, eventive, and social information that can serve as Ground for interpreting a current Figure. A complementary way of looking at the same structures is to ask what they *afford* for the agent during a task (Gibson, 1979). In this sense, context is not only a bundle of available content but also an opportunity structure. Thinking along this, context can be seen as offering indexes to look back to, patterns to re-inspect, and histories that can be re-consulted when the task becomes uncertain. The choice to use context in this way is not always deliberate

¹A detailed methodology can be found in Singh et al. (2025).

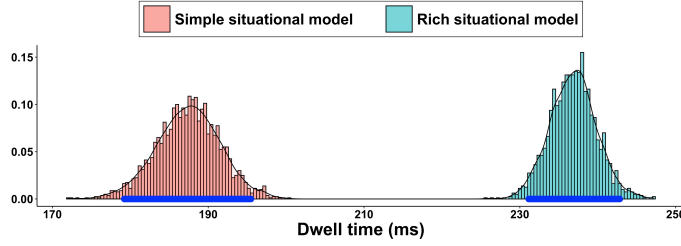


Figure 3.5: Gaze dwell times on the context prior to clarification requests. The histograms display the distribution of dwell times on the physical context for the Simple (left, red) and Rich (right, green) situational models. Participants exhibited significantly longer dwell times in the rich condition, indicating that when the context held diverse information (mixed manners), participants utilized it as an external resource, engaging in extensive "back-referencing" to resolve the under-specified instruction (Singh et al., 2025).

but is ecological. Given limited working memory, it is more efficient to offload parts of the task model into the environment and to rely on the situation itself as a resource, rather than trying to maintain a fully specified internal model at all times (Draschkow et al., 2021; Kirsh & Maglio, 1994). Such a model already exists in the literature, mostly backed by the notion of extended cognition (A. Clark & Chalmers, 1998).

Context, in these terms, does not simply surround the action; in addition, it can be actively exploited as an external memory and as a handle on “what has already happened here” (Hutchins, 1995). The following paragraphs summarize the relevant empirical pattern reported in (Singh et al., 2025), which will then be used to motivate a more general point about how context can function as an opportunity structure for contrastive guidance and repair.

Empirical evidence from Singh et al. (2025).

In the study, participants interacted with a humanoid robot that instructed them to place screw-like objects in front of them. The action’s Path (Where) and Manner (How: vertical vs. horizontal) were manipulated. Over the course of the interaction, these placements gradually built up a situational model on the table, crucially present in front of the participants (Figure 3.4). As explained in the previous section, context was either *simple*, where all preceding screws had been placed with the same manner, or *rich*, where both manners had already appeared before the critical trial. Importantly, instructions then introduced under-informativeness by omitting the manner specification, so that participants had a need to refer back to the context when linguistic information alone was not sufficient. At that point, there was a clear gap between what the robot said and what would be needed to complete the next action. In principle, the action could be performed purely on the basis of an internal task model, without orienting back to the context. However, in practice, participants often first turned back to the physical context, visually scanning the existing screw configuration

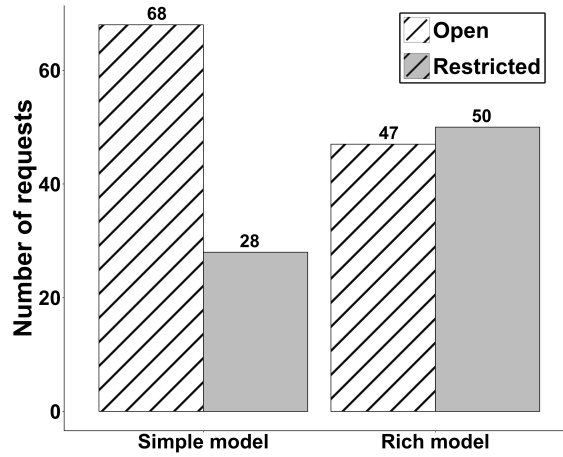


Figure 3.6: Repair types by context. Participants mostly used open repairs (e.g., “How?”) in the Simple context. In the Rich context, where multiple alternatives were visible, restricted repairs (e.g., “Vertical or Horizontal?”) significantly increased, showing that richer contexts support more precise requests.

before initiating a clarification request. This suggested that the situational model was treated as an opportunity to reconstruct and refine the task representation, especially when trouble was encountered.

The exploitation of these external sources was further highlighted when contextual information was pivotal for raising a suitable clarification request. This was shown through the use of repairs and their effect on attention to the context. Within this setting, a distinction was made between open and restricted repairs (Dingemanse et al., 2015). Open repairs (e.g., “How should I put it?”) signal that something is wrong without committing to a particular source of trouble, and hence do not require fine-grained contextual use. Conversely, restricted repairs targeting the precise actions done in the past (e.g., “Should I put it vertical or horizontal?”) specify a missing parameter and therefore presuppose a more fine-grained grasp of the task structure (Schegloff et al., 1977). Crucially, the distribution of repair requests across simple and rich contexts supported the idea that richer contexts offer more opportunities for the fine-grained referencing required for restricted repairs. In rich contexts, restricted repairs occurred relatively more often, so that, when both manners had been provided, participants were more likely to formulate their question in terms of the precise dimension that was under-specified (Figure 3.6). Eye-movement data confirmed this pattern as well: Before producing a restricted repair, attention to the rich context increased, with gaze distributed over the previously established configuration and higher dwell times. Conversely, for open repairs, looks to the context were comparatively sparse. When only one manner was represented, there was little need and gain from exploring the environment, and participants were more likely to raise open request types (Figure 3.7).

In the terms developed throughout this thesis, the same pattern can be re-described

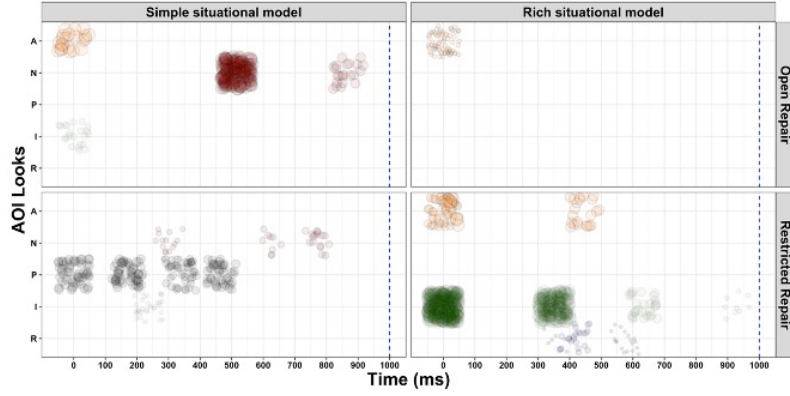


Figure 3.7: Temporal distribution of gaze looks relative to repair onset. The plots show looks to context areas over time for Simple (left) and Rich (right) models. The bottom-right panel highlights that Restricted Repairs in Rich contexts are preceded by dense visual scanning (back-referencing) of the available alternatives, whereas Open repairs (top row) show sparser attention to the context (Singh & Rohlfing, 2025c). The vertical blue line depicts the repair onset.

as showing that context functions as an *opportunity structure* for localizing where repairs can target the source of misunderstanding. Rich situational models do not merely contain more information; they keep multiple alternatives externally available as contrastive candidates, and thereby support restricted repairs that pinpoint the trouble source within the given context. Simple models, by contrast, afford fewer such opportunities and therefore more often lead to open, non-specific queries. In the broader framework of contrastive guidance, this supports the idea that rich contexts are especially facilitative Grounds: they not only provide multiple candidates for being contrasted or negated, but also invite agents to exploit the external representation itself when attempting to understand and repair during interaction.

3.3 Context is what is developed

So far, we have treated context largely as something that is *available*, akin to an agglomeration of situative, functional, locative, eventive, and social information that can be encoded, stored, and retrieved as *Ground* for interpreting a current *Figure*. This perspective naturally aligned with the ContACT-R style model as elaborated in Chapter 2, where canonical Grounds are represented as chunks in memory and relational prominence $R_d(C)$ is computed as a distance between the current “Figural” event and a retrieved Ground. From a phenomenological and interactional point of view, however, this is only half of the story. On this alternative view, as already proposed, context is not first and foremost a container of information, but part of the interactional efforts that participants make (Dourish, 2004). Hence, the question is less *what belongs to*

context in some pre-specified sense, rather more *what is made relevant as context* for a particular activity here and now. Following this, context is occasional since it is occasioned by the activity, the parties involved, and the way participants orient to each other and to the environment (Schegloff, 1992). Seen from this angle, there is no single, static “store” of context that simply sits behind the scene waiting to be referenced. Instead, context *emerges* as people act, speak, and repair each other (Goodwin, 2000). Hence, a restaurant, for example, is not just a fixed backdrop of tables, menus, and bills; rather, it becomes a restaurant “context” for a particular utterance when speaker and recipient align on the canonical script and treat certain elements as relevant for what they are doing, such as ordering, serving, or paying. The same physical setting thus could have acted as a background for an entirely different context, such as a private conversation, a job interview, or a quarrel, crucially depending on which activity and frame the participants enact (Goffman, 1974).

Context, in this interactional sense, is relational not only between objects (Figure–Ground) but also between objects and people, and between moves among participants involved. Let us take the coffee example from the beginning of the chapter. Re-reading the same utterance through a dynamical lens, “Your coffee, sir!” is not simply dropped into a pre-defined context and setting; rather, it helps to conceptualize that context as a shared frame, where the participants’ actions and object states make it a distinct context from others (H. H. Clark, 1996). Likewise, a different utterance in the same context, such as “Today, no tea sir, but coffee”, does more than convey the absence of tea; it highlights the relational aspects. By explicitly invoking the recipient’s usual choice, negating it, and offering an alternative, the speaker actively *develops* the common ground of what counts as normal here, what is exceptional today, and which options are live (H. H. Clark & Brennan, 1991). Hence, the utterance is both about the situation and part of how that situation is made intelligible for both parties. In this sense, contrastive constituents like negation not only make use of prior context; rather, they participate in *making* a rich context by establishing contrast within a context (Dourish, 2004).

This interactional stance also shifts the way we think about relevance and opportunities, as we discussed in the previous section. In the representational view, context is often treated as a potentially huge set of features around an event, with a subsequent selection problem: which of these features matter? (Sperber & Wilson, 1986). From an emergent view, relevance is not decided once and for all by an external observer but is displayed and negotiated in the participants’ own behavior. For instance, a repair initiation (e.g., “Vertical or horizontal?”) shows what part of the prior activity is now being treated as contextually central (here: the manner dimension); similarly, a deictic pointing gesture shows which aspect of the scene is being recruited as Ground for the next move (Goodwin, 2003). What counts as context, then, is precisely what participants mutually orient to as shared resources—what conversation analysis would call the locally achieved “common ground” (H. H. Clark & Brennan, 1991; Schegloff, 1992).

Against this background, the limitations of the current ContACT-R model as conceptualized in Chapter 2 become clear. The ContACT-R account treats context as something that can, in principle, be pre-specified. This pre-specification is already visible, such that for a given context C and dimension d , the model retrieves a Ground $G_d(C)$ from memory based on similarity by computing a distance to the current Figure $F_d(C)$, and then applies a decision rule for using negation or implicit contrast (J. R. Anderson et al., 2004). This is a deliberately static account and therefore does not capture the ideas discussed above, namely, how Grounds themselves are constituted, adjusted, or *co-constructed* in the course of interaction. In other words, common ground cannot simply be written into the model in advance; rather, it is something that participants *co-construct* as they frame an activity, manage breakdowns, and re-describe a situation to resolve misunderstanding (H. H. Clark, 1996; Rohlfing et al., 2026).

Nevertheless, the current ContACT-R model was good at answering questions of the form: given a certain Ground, how contrastive is the current Figure, and under what conditions should we use explicit guidance (negation) rather than implicit verbal guidance? It is much less suited to answer questions of the form: how does that Ground come to be in the first place, how do participants negotiate it turn by turn, which part of the past now counts as relevant history, and how do contrastive moves feed back into what will be treated as context for subsequent actions? Seen from this dynamical view, these emergent processes appears to be missing from purely representational view at this moment (Dourish, 2001).

Crucially, the aim here is not to show that the two stances (representational and interactive) are incompatible. Rather, we want to understand them as highlighting different levels of description. Whereas the representational account may give us a way to formalize contrastiveness by identifying characteristics of a Ground and connecting this to measurable cognitive behavior, the phenomenological and interactional account remind us that Grounds are not just retrieved but are also *developed* through activity building up across turns, negotiated through corrections, *scaffolded*, and sometimes re-framed altogether when an activity shift occurs. In the rest of this chapter, I will use the term *dynamic context* for such interaction-bound aspects.

3.3.1 Context based on dynamicity

One way of bridging the “available” and “developed” perspectives is to think of contexts in terms of their *dynamicity*. At one end, we find relatively static contexts, such as physical layouts, long-standing scripts, and well-practiced affordance structures that change slowly (however, no context is static in a pure sense, hence we might call it “not dynamic”). On the other end, we imagine an interaction, such as a canonical restaurant sequence, as highly dynamic (Dourish, 2004). Whereas contexts that are *not dynamic* can be reasonably approximated by stored templates and fit well with

the representational notion of “what is available”, we find it difficult to explain highly dynamic contexts that are reshaped from move to move, where each utterance and action not only exploits previous structure but also redefines what will count as relevant next (Goodwin, 2000).

For example, in Singh et al. (2025), a series of repairs in a human–robot task can gradually turn an initially vague situation (“put it there”) into a finely articulated local practice with shared labels of paths and manners (action components). In such setups, context is less a pre-existing frame and more a rolling history where an accumulating set of alternatives, expectations, and commitments formed by participants is continuously updated and negotiated (H. H. Clark, 1996). It seems that most real situations combine these two aspects (Section 3.5, Interactional contexts). For example, they combine static elements (material facts, institutional roles) with dynamic ones (emerging task understandings, negotiated semantics, and partner models). Taking an example from Singh and Rohlfing (2024), it was shown that it is possible to hold the task model as relatively less dynamic than the partner model. In this work, it was shown that the variability in the task model can be explained through variability in the partner model.²

In Singh and Rohlfing (2024), on the one hand, the action conceptualization in terms of Path and Manner acted as a relatively stable situational ground rooted in the psycholinguistics literature (Figure 2.4). On the other hand, each instruction-clarification loop following a repair added to or modified what the participants now treated as dynamic aspects, such as the partner’s capabilities to scaffold the task (partner model) (Singh & Rohlfing, 2024). When the partner (here: Robot) was perceived as less capable by the humans—as shown by the inability to provide answers to the asked questions during the course of interaction—the participants showed distributed attention on the partner and took more time to ask questions, showing an “adjustment” to the partner’s capability (Figure 3.8, 3.11).

From the point of view of contrastive contextualism and guidance, both the static and dynamic perspectives matter because they determine *when* the contrastiveness is higher at the task level and for which partner model the explicit contrast will be useful. As we already know, in most static contexts, contrastive guidance can rely on long-standing ground (e.g., negating the routine action). Conversely, in more dynamic settings, explicit contrast may instead work to create a new Ground at the task and partner level. Either way, contrast must not just appear and react to the context, but should also contribute to the development of context by stabilizing certain alternatives as the ones that matter. Although an interactional context with both partner and task models can be treated as a special case where both dynamic and static contents may be found together, observing them in isolation and how they are treated in conventional experiments is worth going through. For an empirical approach, answering the question of what is *not dynamic* (crucially, negating it) makes more sense than defining what is

²For more detailed definitions of task and partner models, refer to Singh and Rohlfing (2024).

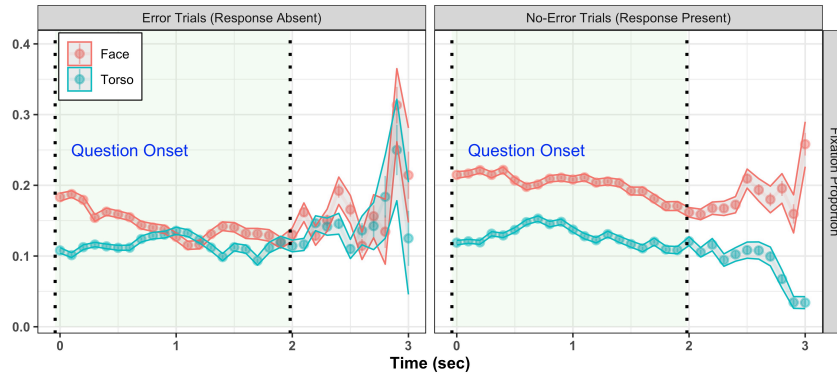


Figure 3.8: Adopted from Singh & Rohlfing, 2024. Looking behavior for two different partner models. Left plot: unsuccessful scaffolding; Right plot: successful scaffolding. Attention on the face and torso was more distinctive where the robot successfully provided a resolution to the repair initiated by the participant as opposed to where the response was absent. The shaded region shows the timeline of repair onset and offset.

static, since there is rarely any static context per se (Dourish, 2004). Hence, the next section turns to this question more directly by taking examples from experiments and reasoning concretely about why they are “not dynamic”. Since dynamicity inherently involves many factors, perhaps negating it provides more light on the background issue.

What is *not* a dynamic setting? Pragmatically, it makes more sense to address the question of “what is *not* dynamic” rather than referring to all the multifarious properties of dynamicity. One way to understand this is to examine the experimental literature and see whether standard paradigms tick all the boxes associated with dynamicity. Since from the beginning we have focused on studying context through the lens of attention, we will take examples from related literature. Most standard paradigms implicitly assume what we can call a *non-dynamic* context. For instance, the Visual World Paradigm (VWP) emerged as one of the prominent methodologies to test the influence of visual context on online language processing (Figure 3.9) (R. M. Cooper, 1974; Tanenhaus et al., 1995). This was perhaps the first time the visual context was brought in lab for controlled investigation. Participants sit in front of a monitor, inspect a picture or grid of objects, and listen to a pre-recorded sentence while their eye movements are tracked.

While these paradigms have been successful in incorporating visual contexts into laboratory studies, they still fall short of capturing the continuous, dynamic nature of social interactions (Huettig et al., 2011). In such setups, a visual or situative layout is presented to the participant, while some linguistic material is played or written, and the task is to process this input while the surrounding scene remains unchanged. Over the course of a trial, there are plenty of temporal dynamics captured at the cognitive

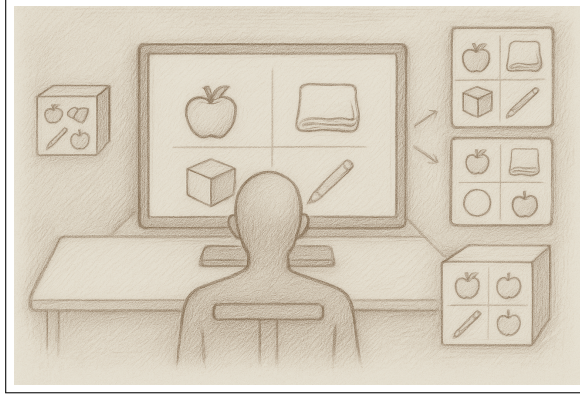


Figure 3.9: A depiction of a Visual World Paradigm setup. Participants are seated in front of a screen wearing an eye-tracker, while stimuli are presented in a trial-by-trial fashion (different trials are shown on the right margin). Cognitive processes are hence measured in a static context as the trial resets.

level—such as lexical access, syntactic parsing, predictions, and attention shifts between objects (Alloppenna et al., 1998; Altmann & Kamide, 1999)—however, the Ground itself does not change. Meaning, the array of objects on the screen, the spatial relations between them, and the set of possible referents are fixed from trial onset until trial offset. The visual context is certainly rich, but it resets on each trial once the sentence ends, and the entire display disappears and is replaced by a new one. Unlike in real-world interaction, participants do not act on the scene (and hence do not actively participate in creating context) in a way that would change it; rather, they act merely as passive observers (Brennan, 2004). Hence, in terms of the contrastive framework, each trial provides a new static Ground, which is a well-defined set of situative and functional features that can be consulted, but not modified, while the sentence (the Figure) unfolds.

Although there is a clear advantage to such setups where static contexts make it possible to time-lock gaze to word onsets and to derive fine-grained curves of how linguistic information incrementally recruits visual information, they lack the dynamics between the partners involved (Henderson & Ferreira, 2004). Moreover, from a modeling perspective, they align naturally with the representational view developed in earlier chapters where, for each context C and dimension d , we can imagine a Ground $G_d(C)$ that is fully specified at trial onset, and thereby ask how relational prominence $R_d(C)$ modulates cognitive processes (J. R. Anderson et al., 2004). In this sense, VWP instantiates the “context is what is available” stance in a particularly lucid way, where the relevant Ground is given once and the only question is how the participant’s internal processes make use of it. However, at the same time, this static regime places a strong restriction on what we can learn about real-world contexts. Since the setting does not provide an opportunity to proactively shape the context, we do not see how context might be

developed across turns, nor how language can feed back into the very structures that will matter for subsequent actions. For contrastive contextualism, a static context can be both a strength and a limitation, such that, on one hand, we gain tight control over which Ground is available, but we lose insight into how Grounds are built up, negotiated, or re-framed through actions.

What is a dynamic setting? An obvious example of a truly dynamic setting would be a real-world interaction. Although such settings provide high ecological validity, they typically yield mostly qualitative explanations of interactive phenomena, which are difficult to measure in any robust sense, especially if one aims to develop a contrastive guidance model on the basis of such findings (Rogers et al., 2011). Hence, we need a setting that offers a systematic pathway to obtain measurable outcomes while still preserving the dynamic character of interaction. Recall that, in dynamic settings, context is not only something that is present at the beginning of a trial; rather, it is something that is *developed* over time as the interaction unfolds. And unlike static paradigms, participants here do not merely observe a scene but act on objects, respond to instructions, and initiate their own moves, where each of these actions can add to, refine, or partially overwrite the existing Ground (H. H. Clark & Brennan, 1991). Put differently, in dynamic settings, the Ground is itself dynamic—it is updated by the very behaviors it is supposed to contextualize.

Hence, to reconcile dynamicity with experimental control, we identified a mid-way solution in one of our experiments, where we treat a Human–Robot Interaction (HRI) setup as a promising pathway. The idea was to maintain experimental control through precisely manipulated robot behaviors while at the same time letting participants interact with the robot as they perform the task and thereby shape a dynamic context (Vollmer et al., 2010). As implemented in Singh et al. (2025), HRI offers a way to *build* a dynamic setting in the lab rather than merely approximating it. The robot can be treated as a semi-autonomous partner whose behavior is tightly scripted at the level of experimental design (e.g., timing, wording, and sequencing of instructions), yet remains open enough for the participant to respond, hesitate, ask for explanations, or even initiate new moves (Rohlfing et al., 2021). Because the robot repeats the same underlying program for every participant, we retain a high degree of control over the data-generating process; at the same time, because the participant is free to act on the environment, to negotiate misunderstandings, and to engage in repairs, the context is constantly re-shaped. Thus, the “dynamic Ground” does not remain a metaphorical notion but something we can measure through controlled experiments, where it is updated by each instruction, action, and repair sequence that enters the mutual task space.³ Crucially, this mid-way solution allows us to move beyond static screen-based

³For a comprehensive description of the methodological implementation and eye-tracking analysis employed in this setup, refer to Singh and Rohlfing (2025b).

paradigms without falling back into purely qualitative fieldwork. Since the robot’s utterances and behaviors are pre-specified, we were able to time-lock participants’ gaze, actions, and verbal responses to well-defined events in the interaction (e.g., the onset of a critical timeline, the omission of information, and the initiation of a repair). At the same time, the unfolding sequence does not “reset” after each trial in the way it does in traditional VWP experiments; rather, the same physical objects remain on the table, but the action history accumulates and changes as the interactional processes take over as the task unfolds. The setting is therefore dynamic both at the cognitive level (as participants update their expectations, hypotheses, and task strategies) and at the interactional level (as the partner model and the shared task history change over time) (Singh & Rohlfing, 2024).

From the perspective of developing a contrastive guidance model, this can be particularly beneficial, since it allows us to study language and vision correspondence in real-time without compromising on interactivity. A model that aims to capture how a tutor uses contrastive relations in context requires not only static snapshots of interpretation but also a record of how the relevant alternatives are brought into, maintained within, and sometimes removed from the Ground as the interaction progresses. In such cases, HRI-based dynamic settings can provide us with a controllable stream of events in which alternatives can be introduced or withheld by the robot, and in which participants’ gaze and repair behavior reveal how the contrast space is being constructed online. By embedding these processes within an HRI framework, we can systematically probe how contrastive cues are deployed and updated in real-time, while still maintaining the replicability and statistical rigour required for formal modeling in a principled way (Scassellati, 2002).

3.4 Interactivity forming its context

As we have treated context from two complementary angles so far: as what is *available* (a structured background that can be encoded and retrieved) and as what is *developed* (a dynamic product of ongoing activity), bringing them together within a methodological framework is necessary. Both perspectives are useful on their own, but they truly come together in what we can call an *interactional context* (Schegloff, 1992). Here, context is neither purely a static representation nor merely a dynamic scene; rather, it is jointly maintained set of resources that interlocutors draw on, orient to, and reshape as they move through an activity (H. H. Clark, 1996). A concrete realization of such interactional contexts—and thereby a its operationalization of the static and dynamic processes therein—can be found in Singh and Rohlfing (2024). Here, the core idea was to identify purely dynamic and relatively less dynamic aspects within a controlled experiment. Since, the task model was based on robust findings in the literature regarding the directionality of Path vs. Manner prominence (offering a relative

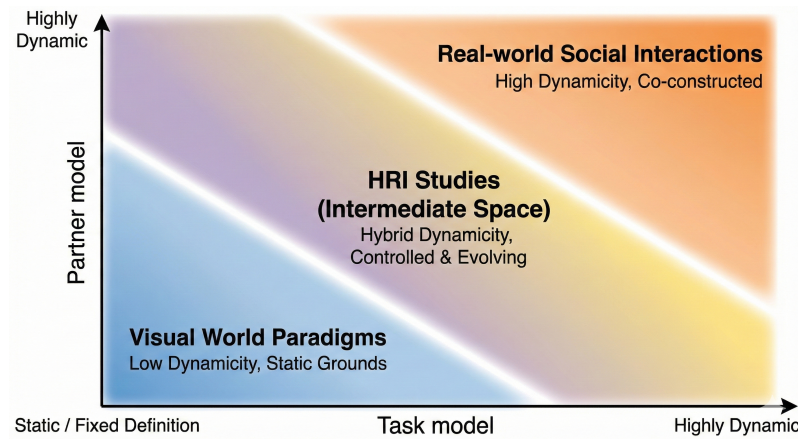


Figure 3.10: The diagram maps different interactive settings based on the variability of their Task Model (x-axis) and Partner Model (y-axis). Traditional Visual World Paradigms (bottom-left) rely on static grounds with low dynamicity, whereas real-world Social Interactions (top-right) involve highly dynamic, co-constructed models. HRI studies occupy the intermediate space, offering a methodological advantage where one dimension can be kept relatively stable to act as a backbone, while the other is allowed to evolve dynamically.

stability to predict cognitive effects a priori), it was treated as a less dynamic aspect of the interaction (Talmy, 2000). Cognitive processes related to the task model were relatively not as dynamic as the processes observed at the social interactive level, such as repairs and turns. In contrast, at the interactive level, the partner model formed during the interaction was highly dynamic, where the roles, responsibilities, and participants' conceptualization of the partner's capability to scaffold the task were difficult to predict a priori and could only be estimated as the interaction unfolded (Brennan, 2004).

Hence, the interactional context, in a true sense, incorporated both dynamic (partner model) and relatively less dynamic (task model) components within a single setup.⁴ In concrete terms, context is now anchored by two components whose contents can be characterized in terms of task and partner models, each changing in its dynamicity (Figure 3.10). If we examine this anchoring closely, an interactional context is never just “everything that happens to be around” but is structured by at least two requirements. First, there must be something relatively stable that can act as a backbone for the activity, such as a task model, a script, or a shared project that constrains what counts as a relevant move (Sebanz et al., 2006). Second, there must be something genuinely dynamic, such as a partner model—a sense of who is doing what and how far one can lean on the other—that allows the same backbone to be inhabited differently across time and across interlocutors (Tomasello, 2008).

⁴Although I treat the task model as static a priori, in a real sense the task model is also dynamic. This distinction is used only to contrast it against the partner model, which is highly dynamic; hence, I use “relatively less dynamic” to situate it against the partner model.

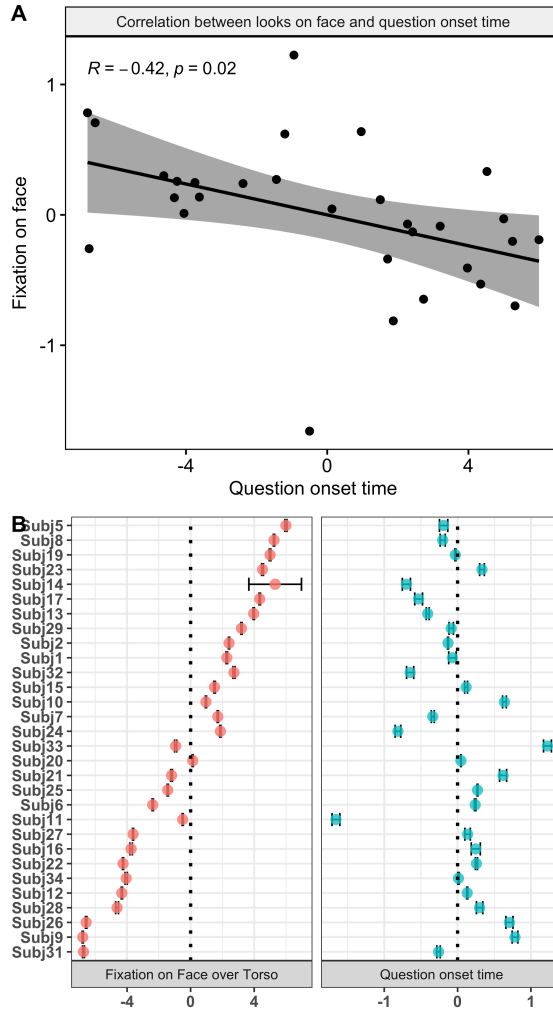


Figure 3.11: The coupling of Partner and Task models. The plots illustrate how the perception of the partner (fixation on the face) influences task performance (question onset time). From Singh and Rohlfing (2024)

In other words, interactional contexts are not only made *richer* by having both parts, rather they entail both dynamic and static components. Seen from this angle, the distinction between “dynamic” and “relatively less dynamic” is not a sharp ontological divide but a functional one for each setup, as shown in Figure 3.10. For instance, task-related expectations—such as that Path information is more critical than Manner for completing a placement action—can be derived from robust findings and therefore predicted to some degree a priori (Talmy, 1985a). They still change, but they do so slowly and under stronger constraints. Conversely, partner-related expectations are typically formed on the fly; for example, whether the robot (or human tutor, co-driver, or teammate) is seen as a capable scaffolding partner depends on how they respond in very local moments and whether they repair a misunderstanding (Epstein et al., 2017). These expectations can shift within a few turns of talk. Thus, in an interactional

context, the relatively less dynamic part keeps the activity recognizable as the same task, whereas the more dynamic part keeps it negotiable and responsive to the moment.

Importantly, this layered structure is not unique to the HRI setting discussed above. In human–human dialogical setup, for instance, the curriculum or problem set functions as a relatively less dynamic constraint, while the teacher’s emerging sense of what the learner currently understands and the learner’s evolving judgment of how much they can rely on the teacher form a highly dynamic partner model (Chi et al., 2001). In everyday joint activities such as planning a route, the physical layout of the environment and the goal provide a relatively stable task model, whereas the ongoing negotiation about who leads, who checks the map, and how task responsibility is distributed when things go wrong is dynamic and often only visible in micro-moments of gaze, hesitations, and repairs (Pickering & Garrod, 2004). In both cases, we would hesitate to call the situation truly interactive if either side were missing, if the task kept changing arbitrarily, or if the partner’s role was fixed and unresponsive.

The HRI setting makes these two requirements especially transparent because we can design them separately (Singh & Rohlfing, 2024). As can be seen from Figure 3.10, either dimension—the partner or the task model—can be kept relatively less dynamic, such that the other can be controlled and investigated. We can, for example, hold the task model constant by scripting the same distribution of Path and Manner information across trials, while letting the partner model evolve as participants probe the robot’s capabilities, sometimes encountering successful scaffolding, sometimes not. But the conceptual point is more general: whenever interaction unfolds in time, the context that is doing the real explanatory work is an *interactional* one that couples a relatively stable organization of the task with a dynamic organization of the partner. The methodological challenge, then, is to construct settings in which these two components (task and partner models) are both present and can be analytically separated.

A relatively less dynamic backbone that can be characterized in advance (for example, in terms of task affordances or known asymmetries such as Path over Manner), and the dynamic layer in which partners negotiate, adjust, and sometimes contest how this backbone is to be used here and now. Only then can we speak of an interactional context in the stronger sense that matters for contrastive guidance and for modelling how interlocutors orient to, and re-shape, their shared Ground over time.

Finally, the dimensions should be treated as independently but closely related; only by looking at both layers together can we make sense of the interactive context. For example, in Singh and Rohlfing (2024), it was found that the partner model conceptualization affected the task model, in the sense that the variance in the task model was partly explained by the variance in the partner model. In more concrete terms, when participants perceived the robot to be less capable of providing resolution, the repair onset was delayed (Figure 3.11) (Singh & Rohlfing, 2024).

3.4.1 Discussion: why context matters for contrastive guidance

This chapter has asked a seemingly simple but central question: what is meant by “context”, and how should it be treated as a basic building block for contrastive guidance? The reason for raising the question at this point is methodological as well as conceptual. Across the work brought together in this thesis, contrastive guidance repeatedly relies on a Figure–Ground relation: a current action (Figure) is interpreted against a structured background (Ground) that supplies alternatives, expectations, and constraints. Without a more explicit notion of context, the term risks remaining a ill-defined concept, and it becomes difficult to compare the different ways in which context is operationalized across the modelling and empirical parts of the thesis.

A first way to approach context in the literature treats it as something that is already *available*, akin to structured information that can be retrieved when needed within interaction (Dourish, 2004; Minsky, 1975). From this perspective, the guiding intuition has been that context is not an ill-formed collection; rather, it has a layered structure that can be specified across dimensions such as situative, functional, locative, eventive, and social. These dimensions jointly furnish the Ground against which a current Figure (or action) is interpreted. Along this line, the symbolic ACT-R account (the ContACT-R model) from Chapter 2 fits comfortably into the picture: for a given context C and dimension d , a set of Ground representations $G_d(C)$ can be pre-defined and retrieved whenever necessary to derive relational prominence $R_d(C)$, which tracks how far the current Figure stands from these Grounds (J. R. Anderson et al., 2004). On the empirical side, the HRI experiment made it possible to operationalize context along informational dimensions—for instance, the distinction between rich and simple situational models (Singh et al., 2025). Rich contexts contained multiple relevant alternatives as part of the Ground, whereas simple contexts contained fewer such alternatives, and this difference was reflected in how participants formulated repair requests.

Once this representational view is adopted, its limitations also become visible. If context is simply everything that is available, it becomes overwhelmingly large, such that almost anything in the vicinity of an event could be listed as background while most of it never enters the interpretative scope. This raises a selection problem about which parts of this potential context are actually relevant? Although the ContACT-R formulation overcome this challenge through cue-based retrieval, it still presupposes a large store that has to be managed (Lebiere & Anderson, 1998). Moreover, the representational view risks treating context as static, as if it were a set of features that can be enumerated independently of what participants are actually doing (Suchman, 1987). These shortcomings motivate the second approach in the literature, where context is treated as something that is *developed* during interaction (Rohlfing et al., 2026).

Under this approach, context is *co-constructed* and reshaped as people act, speak, and repair one another over time (Goodwin, 2000; Rohlfing et al., 2020). From this

perspective, context is not merely a container of information behind the scenes; it is actively constituted through joint activity, and what counts as context for an utterance is whatever participants currently orient to as relevant (Schegloff, 1992). This can empirically be shown by online repairs, deictic gestures, and shifts in gaze reflecting as a contribution to creating a more dynamic context by displaying which part of the prior activity has become central. An example was taken from Singh and Rohlfing (2024) within HRI setup, where restricted requests such as “Should I place it vertically or horizontally?” make the manner dimension of the action history relevant, whereas open requests such as “How should I put it?” leave the trouble source under-specified, and this difference is mirrored in participants’ gaze behaviour (Singh & Rohlfing, 2025b). Similarly, on the social dimension, contrastive utterances such as Today, no tea, sir, but coffee do not simply rely on a usual restaurant script, rather they reshape common ground by marking what is normal and what is exceptional for the present relationship (H. H. Clark, 1996; Levinson, 1983). In this sense, negation and other contrastive devices not only exploit existing context but can also contribute to making context more specific for a situation.

At this point, it is important to avoid an overstatement. The claim is not that a satisfactory account of interactive context always requires both a representational and an interactional approach, nor that purely qualitative work cannot investigate interactive context on its own terms. Rather, for the goals of this thesis and in particular, for experimentally tractable accounts that connect formalization with observable interactional behaviour, it is useful to keep both perspectives in view and to make their interface explicit. This interface is captured here under the specific notion of *interactional context* which is not a third type of context on top of available and developed, but the analytic layer at which relatively stable structures (what can be treated as available Ground content) and unfolding interaction histories (what becomes relevant through activity) are brought into relation. In other words, interaction provides the mechanism by which parts of what is available become selected, updated, and locally ratified as the Ground for interpreting the next Figure.

This framing also clarifies the role of the dynamicity axis introduced in Figure 3.10. Here the axis was not intended to reify a strict static vs dynamic dichotomy rather it offered a way to characterize how quickly different contextual structures change on the timescale of an interaction. Structures that are stable for the duration of an experiment can be approximated as available Grounds and thus align with a representational treatment. By contrast, partner models, emerging repair practices, and accumulating action histories evolve as interaction unfolds and cannot be fully pre-defined. The value of this approach was particularly highlighted for the HRI settings where it makes both layers observable: relatively stable task structure can be controlled, while genuinely interaction-shaped context can be tracked through gaze, action, and repair.

Evaluated against this background, the current ContACT-R formulation is well suited

to questions such as how contrastive is the current Figure against a Ground? However, it is much less suited to questions such as how a Ground comes into existence at all, or how it changes as participants negotiate meaning. This is not a complete failure of the model rather it is a reminder that a full account of contrastive guidance, as it occurs in interaction, must also address the processes by which contextual relevance is established and maintained.

Finally, this motivates the organization of the remaining chapters. The next chapters on scaffolding take up the interactional perspective more directly by analysing how agents establish relevance, monitor understanding, and use contrastive resources (e.g., negation, repair, and other targeting moves) to shape what counts as Ground in the moment. In this way, the thesis connects the papers' contributions into a single line of argument. First, formalizing contrast in terms of Figure–Ground structure provides a basis for specifying *what* can be available as context. Second, the interactional analyses and the HRI paradigm provide a method for studying *how* such Grounds are selected, reshaped, and co-constructed in real time.

Chapter 4

Interaction and Scaffolding

4.1 Studying interaction

When we envision an interactive scenario, we tend to picture a dialogical setting where the agents involved (be they natural or artificial) engage in a turn-taking protocol—or what is often termed conversational turn-taking (Sacks et al., 1974). Hence, an “interaction” is typically defined by a setting where, at minimum, a turn-taking process is taking place. Crucially, this turn-taking is considered essential for effective coordination, the development of joint understanding, and cooperative social interaction (Levinson, 2016). Research into the processes involved in turn-taking as a central problem in social interaction dates back to the seminal work of Sacks et al. (1974). Since then, studying interaction has essentially meant studying the systematics of the organization of conversational turn-taking, such as fundamental features, repairs, and the specific time windows between turns and overlaps. These studies predominantly aim to derive a mechanistic explanation for turn-taking phenomena, often relying on a *turn-construction component* to identify points where a speaker transition is possible, and a *turn-allocation component* with rules for selecting the next speaker, aiming to minimize overlap and silence (for a recent review, see Cosper (2025)). Consequently, the question of interaction is often boiled down to solving turn-timings and explaining them mechanistically (Skanze, 2021). Using large volumes of conversational data to find the point in an utterance where the turn is most likely to happen has been of significant practical achievement, especially for real-world implications such as HRI.

However, even if large-scale, data-driven methods have addressed turn-taking in Human–Robot Interaction (HRI) to a considerable extent, an open issue remains: interaction with artificial agents is often experienced as phenomenologically distinct from human–human exchange. Meaning, a focused investigation on turn allocation and turn-timing captures an important component of interaction *inter alia*, but it does not exhaust the interactional processes that sustain natural dialogue. In particular, differences in how contributions are monitored, grounded, repaired, and adapted *in situ* provide plausible explanations for why an interaction can be temporally well-formed yet still appear artificial (De Jaegher et al., 2010).

This point can be sharpened by separating the organization of turns from the organization of a joint activity. Systems can alternate contributions with minimal gaps and overlaps, while still providing limited evidence of partner-oriented monitoring, an idea previously proposed as participatory sense-making (De Jaegher & Di Paolo, 2007; De Jaegher et al., 2010). Human interaction, by contrast, is routinely organized around maintaining mutual orientation to the task and to the partner’s evolving understanding (Rohlfing et al., 2021). In this sense, turn-taking is better treated as an observable surface structure through which deeper coordination mechanisms become visible, rather than as a complete account of interaction.

One potential way to characterize these mechanisms is through *scaffolding*: temporally extended, adaptive support by a more knowledgeable partner that enables a less knowledgeable partner to progress in a task (Wood et al., 1976). Crucially, scaffolding presupposes that the tutor continually tracks how the partner is doing and adjusts support accordingly; this partner-oriented tracking is commonly discussed as monitoring within accounts of grounding and coordination (Rohlfing et al., 2021). Recent work in sXAI likewise emphasizes scaffolding and monitoring as intertwined interactional requirements, rather than optional route to turn management (Rohlfing et al., 2026).

This shift in focus also reiterates the methodological issue developed in the previous chapter: whether context should be treated primarily as something that is already *available* to cognition, or as something that is *developed* through joint activity (Dourish, 2004). Because scaffolding depends both on relatively stable resources (task constraints, canonical action patterns, established grounds) and on context that emerges online through questions, repairs, gaze, and other coordinating moves, it helps bridge the gap in how dynamical context can be conceptualized and modeled—as both an interpretive resource and an interactional product.

Against this background, the aim of the current chapter is to treat scaffolding as a central explanatory construct for interaction, and to use HRI as a methodological bridge between controlled experimentation and interactional richness. As highlighted in the previous chapter, HRI makes it possible to retain a controlled task, a well-defined workspace, and targeted manipulations, while also allowing partners to coordinate over time through action sequences, clarification requests, and responsive guidance (Rohlfing et al., 2021; Singh & Rohlfing, 2025b). In this setting, the present thesis moves beyond a static notion of Grounds toward an interactive account in which repair and monitoring can be operationalized and measured. Concretely, this motivates extending the previous ContACT-R model into an interactive framework (InterACT-R). Rather than treating Grounds as static templates retrieved at the start of a trial, the extended model must accommodate dynamic context as it is reshaped by the learner’s actions, shifts in what is treated as relevant, and the explainer’s scaffolding moves over time.

The remainder of this chapter proceeds in three steps. First, it briefly characterizes how traditional and interactionist approaches study language, and highlights the limitations

of each if taken in isolation. Second, it uses the HRI paradigm developed in Singh and Rohlfing (2025b) to show how experimental control and interactional processes can be combined in a single design, with particular attention to monitoring and repair as mechanisms that support scaffolding. Third, I sketch how ContACT-R can be upgraded toward InterACT-R, such that contrastive guidance is no longer applied to a fixed scene but is adapted to a context that is continuously shaped by the unfolding interaction.

4.1.1 The traditionalists

The first side, which we have been loosely calling the traditionalist or cognitivist theory, starts from the idea that language processing can be treated as an internal computational problem (Fodor, 1975). On this view, the mind is a system that operates on structured representations, and context is part of the input to that system (Pylyshyn, 1984). Context can therefore be encoded as features, stored in memory, and retrieved when needed to support prediction, ambiguity resolution, or the updating of a situation model (Zwaan & Radvansky, 1998). Much of mainstream psycholinguistics methodology sits comfortably in this tradition. For instance, in classic sentence processing within Visual World Paradigms, participants see a controlled display and hear a pre-recorded sentence while their eye movements are tracked over time (Tanenhaus et al., 1995). The crucial point is that the display itself is fixed during the trial where nothing about the visual Ground changes as the sentence unfolds, except the participant’s internal cognitive state. This fits very well with a modular or “massive modularity” picture within cognitive science, where language, vision, and action are treated as separate components that pass information between them but can still be studied in relative isolation (Carruthers, 2006; Fodor, 1983). Within such paradigms, context is operationalized as static and cleanly bounded. Hence, context is “available” when the trial starts, not something that is reshaped by what the participant does. The ContACT-R model in Chapter 2 belongs to this family by assuming that, for a given context C , a set of Ground representations $G_d(C)$ is present in memory and can be retrieved by the cognitive model (J. R. Anderson, 2007). Relational prominence $R_d(C)$, in this sense, is defined as a function of how the current Figure matches or deviates from these Grounds, and a decision rule then determines whether implicit cues are enough or whether explicit contrast (for example, negation) should be used.

In doing so, the model inherits many virtues of the traditionalist stance. For example, it forces us to be explicit about what is represented, how retrieval works, and which parameters control the use of contrastive guidance. It also allows us to connect these internal states to fine-grained behavioural measures such as eye movements and recall. At the same time, this stance comes with obvious limitations. Since context is fixed within a trial, there is no room for the participant to change it by acting on the world, and there is no partner who might co-construct. Hence, the rich interdependencies

between language, vision, and action in real interaction are partly lost in a one-way mapping from input to response. Recent work by Singh and Rohlfing (2025b), arguing for “massive interactivity” in cognitive science, explicitly calls for frameworks in which language, vision, and action are treated as mutually shaping processes unfolding in real time rather than as separate processes in a pipeline (Spivey, 2007). The traditionalist stance in this work should therefore be seen as offering one important piece of the puzzle: a way of formalizing how a single agent might behave given a particular context, but not yet an account of how that context itself is built up and negotiated in interaction.

4.1.2 The interactionists

The second stance, what will be referred to here as *interactionist*, begins from almost the opposite approach compared to traditionalists. It takes as its starting point a longer interactionist tradition spanning ethnomethodology and conversation analysis, in which language is treated as something that people *do together in time*, rather than something that is passively processed by isolated individuals (Garfinkel, 1967; Sacks et al., 1974; Schegloff, 1992). Within this landscape, H. H. Clark (1996) provides a useful cognitive-pragmatic synthesis of how such coordination is achieved in joint action. On this view, an utterance is always a move in an ongoing activity. It is addressed to a partner, produced at a particular moment in a shared history, and it makes sense only against the background of that history (Schegloff, 1992). Context in this sense is not just a set of features in a visual scene or a bundle of stored representations in memory; rather, it is what participants treat as relevant for the task at hand, and this relevance is displayed and negotiated turn by turn (Goodwin, 2000).

Crucially, in such a framework, phenomena like repair and scaffolding are central. Interactive repairs, for instance, are not merely exceptions or errors to be discarded before analysis, but are rules of interaction (Albert & de Ruiter, 2018; Schegloff et al., 1977). Thus, conversational breakdowns are structured practices that reveal how interlocutors conceptualize the source of trouble and how they allocate responsibility for resolving it (Dingemanse et al., 2015). Taking an example from Singh and Rohlfing (2024), open repairs such as “How?” or “What?” signal that something has gone wrong but leave the locus of misunderstanding unspecified. Restricted repairs such as “Vertical or horizontal?” target a specific missing source and therefore presuppose a more fine-grained grasp of the evolving context. In this way, the form of repair already tells us something about the task model that the speaker has constructed and about which aspects of the shared history they are currently treating as Ground (Singh et al., 2025).

The interactive psycholinguistic framework as proposed in Singh and Rohlfing (2025b) is motivated by this line of thinking. In that setup, unlike in static paradigms where participants view a fixed display while listening to sentences, participants interact with a humanoid robot that instructs them to carry out a sequence of object-directed actions

in a shared workspace. Because the context is now dynamic, we can observe how language, vision, and action influence each other in ways that are simply not visible in traditional static paradigms, such as Visual World Paradigms (Tanenhaus et al., 1995). Our setup, for example, enabled us to investigate questions that are relevant for interactionists, such as: a) the choice of repair type modulates visual attention, and this modulation depends on how informative the current context is, which can be measured as the interaction unfolds; b) restricted repairs are more frequent in rich contexts and are preceded by longer dwell times and more distributed scan paths over the action model, suggesting that participants use the evolving scene as a resource to locate and target the exact trouble source; and c) open repairs are more common in simple contexts and are associated with sparser looks to the context (Singh & Rohlfing, 2025b). Here, context is treated as something that is being developed, and language is one of the mechanisms by which this development is made visible. Despite the challenges that dynamic context poses for an interactionist account of language–vision processing due to bidirectional couplings among language, vision, and action, such processes remain tractable in experimental setups that preserve control while allowing participants to act on the world (S. E. Anderson et al., 2011; Brennan, 2004).

As mentioned earlier, in this chapter, our aim is to not abandon the traditionalist stance in favor of the interactionist one, instead, we take them as two sides that need to be held together if we want a theory of contrastive guidance that is both formally precise and ecologically grounded. The interactive HRI paradigm is an attempt in this direction, to retain enough control and support modeling for the InterACT-R, but it also allows context to change and lets participants display their understanding through their actions and repairs. In the following sections, I will use this setting as an example to sketch how language and visual processes can be studied using Human-Robot Interactional setups. Further, I will elaborate on one of the central mechanisms of interaction, i.e., scaffolding, and how the previous model (ContACT-R) can be extended toward an interactive scaffolding architecture (Wood et al., 1976), in which contrastive guidance is not computed solely for static Grounds but is adapted to, and helps shape, a developing interactional context.

4.2 HRI as a field to study language and visual processes

One of the reasons why the field of Human-Robot Interaction (HRI) feels “interactive” yet simultaneously representational is that it is often viewed as a promising domain to implement interaction, while at its core, the essence of true interactivity sometimes feels missing. There is something midway about HRI setups in sense that one cannot affirm that it is not interactional, but at the same time, one cannot deny it is often purely representational. This uncertainty stems from research on anthropomorphism, where the duality of robots being inanimate yet treated as social partners, remains an

unresolved debate on their ontological status (Epley et al., 2007; Kahn Jr et al., 2006). Thus, HRI offers a middle ground between traditional laboratory paradigms with static contexts and fully naturalistic human–human interaction, where many potentially relevant variables are difficult to control. A natural question at this point is that whether this middle ground can be tested in a way that respects both the need for experimental control and the inherently interactive nature of language? In the past, HRI has been shown to be particularly promising field for this purpose (Sciutti et al., 2012). Here, the experimenter can design a task in which a robot provides instructions and responds, while a human participant acts on objects in the physical world. This setup already brings in a key ingredient that is missing from static paradigms: *dynamic context*. The workspace changes as both agents act, where a sequence of instructions and object manipulations gradually builds up a shared situational model (Rohlfing et al., 2003). At the same time, the robot’s behavior can be scripted or parameterized in such a way that specific linguistic or visual stimuli are systematically manipulated. In other words, HRI allows us to have it both ways: we can keep enough structure to ask traditional psycholinguistic questions about timing and processing, and we can embed these questions into an interactive scenario in which context is both available and developed. The current thesis also capitalizes upon Singh and Rohlfing (2025b) to show that it is possible to exploit this HRI setup to ask traditional questions in an interactive framework. For instance, participants interacted with a humanoid robot that instructed them to place screw-like objects in front of them, varying in action Path and Manner. Over time, this gave rise to different kinds of situational models on the table: some with simple histories (only one manner configuration), others with rich histories (both manner alternatives are revealed). The robot could also deliberately produce under-informative instructions, for example, by omitting the manner. In such moments, language, vision, and action became tightly coupled; participants could inspect the evolving context, decide whether they understood enough to act, or initiate a repair to ask for clarification (Dingemanse et al., 2015). All of these behaviors were informative about how language and visual processes operate in a dynamic context.

Motivated by this approach and its potential, we see HRI not primarily as a field of application on its own, but rather as a field for *studying* language and visual processes under interactive conditions. The key idea is that the robot can serve as a controlled, reliable partner whose contributions are held constant across participants, while the human side is free to show the natural variability of interaction, such as different repair strategies, different uses of gaze, and different ways of exploiting the context (Vollmer et al., 2010). This makes it possible to frame empirical questions in a way that is familiar from traditional psycholinguistics, but to answer them using measures that are meaningful from an interactionist point of view.

As far as psycholinguistics and its methods are concerned, the focus on *language* and *vision* is deliberate. On one hand, repairs make visible how interlocutors reveal language

processes online while negotiating understanding. On the other hand, gaze simultaneously provides a complementary trace of what is attended to and treated as relevant as the shared context develops. Together, these modalities allow a more complete characterization of interactional context than either modality in isolation, and they also make it possible to connect the controlled structure of an experimental task with the real-time characterization of coordination.

Hence, I first examine HRI as a methodology for studying *language processes*, with particular attention to repair during a robot’s guidance. This goes beyond the empirical claims in the individual papers by using repair not only as a dependent measure, but as an analytic lens for how scaffolding is accomplished through sequential organization. Second, I examine HRI as a site for studying *visual processes*, focusing on gaze in dynamic context: not only as an input channel for reactive perception (Admoni & Scassellati, 2017), but also as a proactive resource through which participants structure the interaction and display orientation to what is currently relevant (S. E. Anderson et al., 2011).

4.2.1 Language processes

From the perspective of the current work, the most obvious way in which HRI gives us access to language processes is through *interactive repair* (Albert & de Ruiter, 2018; Dingemanse et al., 2015). In a Human-Robot explanatory dialog, when the robot’s guidance is under-informative or ambiguous, participants can either take a guess, hesitate, or turn back to the robot to initiate a repair (Singh & Rohlfing, 2024). This is enabled by the interactional flexibility of the setup where participants can deploy repair moves as needed, rather than responding to experimenter-imposed stimuli, which in-turn naturally reveals linguistic processes due to spontaneous adjustments that participants bring in to keep the interaction going. As such, they also inform us about the task and partner model (Singh & Rohlfing, 2024), which aspects of the context are currently treated as Ground (Singh et al., 2025), and how contrastive guidance is being requested or negotiated (Singh & Rohlfing, 2025b).

For example, in the interactive experiment above, we systematically manipulated the informativeness of the robot’s instructions (Singh & Rohlfing, 2025b). On some trials, the robot specified both Path and Manner, making the required action explicit. On critical trials, however, one element was left unspecified, typically the Manner (“Put the red object on A” without specifying *How* it should be placed). Here, participants found themselves in a situation where linguistic input alone was insufficient to fully determine the next action. Following interactionist stance, we did not treat this as a failure of experimental design, but as an opportunity to observe how participants use repair, and measure online language mediated visual attention (Singh & Rohlfing, 2025b).

Firstly, two broad types of repair as known in the literature emerged. **Open repairs**, such as “How should I put it?” or “How?”, elicited a problem but did not specify its locus (Dingemanse et al., 2015). **Restricted repairs**, such as “Should I put it vertical or horizontal?”, explicitly mentioned and targeted the source of trouble (Kendrick, 2015). This distinction had a natural interpretation within the contrastive framework. Open repairs signal that the speaker senses a mismatch between their task model and the robot’s guidance, but they have not (or do not wish to) anchor this mismatch in a particular Figure–Ground relation. Restricted repairs, by contrast, treat a specific dimension (here, Manner) as the relevant Ground and pose a direct contrastive question about it. In this way, restricted repairs are requests for contrastive guidance, since they ask the partner (robot) to provide the kind of information that contrasts with the Ground, which we have treated as central to explicit contrast in earlier chapters.

Secondly, the distribution of these repair types across simple and rich contexts is also informative about the use of repairs in specific contexts as they evolve. In **simple contexts**, where only one manner was revealed, open repairs were more frequent. This suggests that participants felt a general sense of under-informativeness without necessarily treating both manners as equally live alternatives (Singh & Rohlfing, 2025b). In **rich contexts**, where both the manner possibilities had already been established in the interaction, restricted repairs occurred relatively more often. Here, participants seemed more willing to frame the problem explicitly as a choice between two known options. The finding fits well with the broader claim that rich contexts support more fine-grained linguistic encoding, such that when the Ground encodes multiple relevant alternatives, it becomes easier to articulate which relation is in question.

Beyond repairs, the HRI setting can also shed light on how people interpret contrastive guidance in incoming instructions. For example, in Groß et al. (2023), when the robot used negation (“Don’t press—turn”), participants could treat this not only as a description of how the action differs from a canonical pattern but also as a scaffolding move that corrects a prepotent routine. The same utterance can thus serve both as a contrastive encoding of a functional relation and as a teaching move that reshapes the participant’s task model (Wood et al., 1976). Hence, scaffolding seems to be an integral dimension of interaction, and such uses of language are difficult to study in purely static paradigms because they require a history of interaction in which routines, alternatives, and misunderstandings have had a chance to emerge.

Going further, HRI can provide a setup in which language processes can be studied as they operate *within* interaction, rather than only as isolated responses to stimuli. In this setup, repairs and responses to contrastive guidance can act as a proxy to investigate how participants actively manage gaps in their understanding, how they recruit contextual alternatives as Grounds, and how they react to partner’s scaffolding moves (Rohlfing et al., 2021). These measurements are needed to investigate and develop an interactively informed model (InterACT-R) as an extension of static models.

4.2.2 Visual processes

If language tells us how participants *talk* their way through the task, eye movements tell us how they *attend* their way through it. In static paradigms, gaze is typically treated as a window onto moment-to-moment processing, where looks to objects are taken as indices of referential activation or prediction (Tanenhaus et al., 1995). In the HRI setting, gaze certainly plays this role as well, but it also does more. It serves as a means of monitoring the partner, checking the outcome of actions (H. H. Clark & Brennan, 1991), exploiting the environment as external memory (A. Clark & Chalmers, 1998), and structuring the repair process (Admoni & Scassellati, 2017).

One basic function of gaze in our HRI tasks is to keep track of the evolving situation. As participants perform actions according to the robot’s instructions, they see and encode the events as it develops and change. This visual record is not merely to create a superficial pattern; rather, it embodies the history that, in earlier chapters, we have described as part of the *eventive* dimension. When the robot later produces an under-informative instruction, participants often direct their gaze back to this history before deciding how to linguistically respond (Singh & Rohlfing, 2025b). For example, in rich contexts, where both manners are present, this “referring back” is more pronounced; scan paths become more distributed and dwell times on the history increase in the time window, especially before a restricted repair. This suggests that participants are visually inspecting the context to reconstruct what possibilities have been established so far and to identify which element is missing now. This can be considered the cognitive function of gaze, in addition to the communicative function.

For example, gaze also plays an important role in how participants coordinate with the robot. Although the robot’s gaze behavior is typically scripted, participants still monitor its face and body (Singh & Rohlfing, 2024), especially to pick up cues about turn-taking and task progression (Emery, 2000). Hence, looks toward the robot can signal readiness to receive the next instruction, or can accompany a repair initiation, thereby revealing the verbal behaviour as a bid for joint attention to the problem (see Figure 4.1). In this way, visual attention is not only about objects and layouts, rather, it can also inform us about the social dimension of the interaction (Skantze, 2021).

A further function of gaze in a dynamic context is to offload memory and support simulation. When participants rehearse possible actions in their mind, they often do so by visually tracing potential paths or by briefly fixating on candidate positions (Kirsh & Maglio, 1994). In the HRI study, we see this in moments where the robot has given an instruction that is informationally sufficient, but the participant still spends time looking over potential target locations before acting.

These patterns are informative because they reveal how participants are exploring the space of options that the context affords, even when the linguistic input is complete. This phenomena has been attributed to the working memory utilizing the external

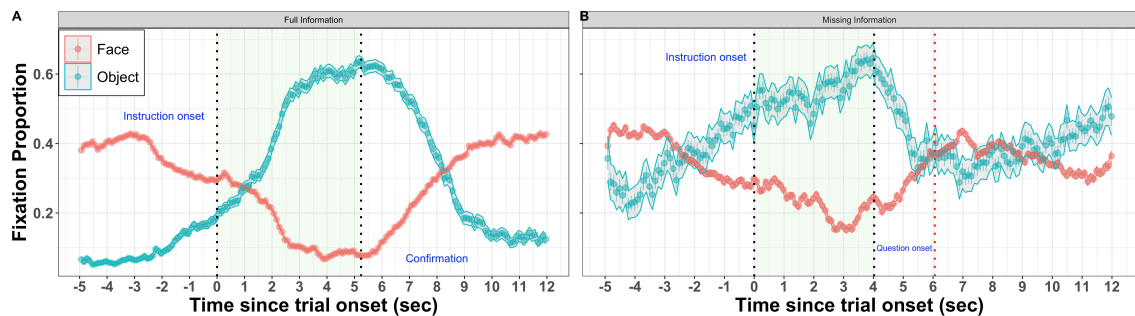


Figure 4.1: Temporal dynamics of social gaze during instruction (Singh & Rohlfing, 2024). The plots display the proportion of fixations on the Robot’s Face (red) and Task Objects (teal) over time. (A) Full Information: A routine pattern where gaze shifts to the object to execute the task and returns to the face for confirmation. (B) Missing Information: When the robot’s guidance is under-informative, participants rapidly shift their gaze back to the robot’s face (approx. 4s), visually signaling trouble and bidding for attention before the onset of their verbal repair question (red dotted line).

resources for ecological reasons (van Ede & Nobre, 2023). Contrastive guidance can interact with this process by highlighting some options and suppressing others, thereby shaping the visual inspection patterns that precede action. Crucially, this aspect can be very important for scaffolding, where the internalization of the task can be treated as comprehension, based on which scaffolding can be more or less provided depending on how the user relies on internal or external sources, an aspect we look at in the upcoming section (Ballard et al., 1997).

Finally, gaze provides a way to see how static and dynamic elements of context are being used together (section 3.3). Some aspects of the workspace are stable across trials (the layout of positions, the identity of objects), while others accumulate through action (which positions have been used, which manners have appeared). Eye movements show how participants navigate between these layers. Early in the interaction, looks may be more dominated by the static geometry of the scene, which later, as a history accumulates, increasingly reflects the dynamic structure of that history (Singh & Rohlfing, 2024). For an interactive model of contrastive guidance, this suggests that visual processing cannot be reduced to a simple mapping from words to fixed referents. It must be sensitive to how the Ground itself is changing and to how that changing Ground is being used as a resource for understanding and repair.

In sum, HRI gives us access to cognitive processes that are intimately tied to interactional context. Here, gaze is not only an index of comprehension, but also a tool for exploiting external resources and signaling engagement. Together with the language-based measures discussed above, this makes HRI a rich field for studying how contrastive guidance may operate in dynamic, shared environments, and substantiates the development of a scaffolding model that takes into account visual and linguistic processes.

4.3 Scaffolding

Up to this point, contrastive guidance has mostly been treated as a way of shaping how an observer encodes an event, for instance by using negation and related contrastive devices to raise relational prominence, re-balance goal bias, and make certain Figure–Ground relations more salient (Singh & Rohlfing, 2025a). In experimental terms, this type of guidance largely assumed a static state with one-shot guidance selection.

However, outside tightly controlled trials, guidance is rarely unidirectional. In natural tutoring and instruction, it is better characterized as tailored support provided by a more informed partner to a less informed partner in order to enable successful task completion (Wood et al., 1976). This support is recursive, which unfolds across turns, is adjusted in response to the learner’s observable behavior, and relies on the coordinated use of context and prior knowledge as part of the guidance itself (Rohlfing et al., 2021). In that sense, as highlighted earlier, scaffolding is not merely an “add-on” to turn-taking, rather it exists for why certain turn sequences take the form they do when an interaction is goal-directed.

A large part of HRI research has understandably concentrated on coordination problems such as turn allocation, timing, and smoothness. Yet these measures alone do not specify the functional organization that makes interaction pedagogically effective. For interactions involving explanations in particular, the central question is how a tutor monitors the partner’s progress, detects when support is needed, and reshapes guidance so that it improves future performance rather than merely accounting for the present move (Buschmeier et al., 2025; Chi, 2009; Lajoie, 2005). Within the scaffolding tradition, this adaptive organization is commonly described against the background of the learner’s changing abilities and the Zone of Proximal Development (Vygotsky, 1978). When the learner can only act successfully with support, guidance is comparatively explicit, and as competence develops, effective tutoring typically reduces that support in a systematic way, often discussed under the label of *fading* (Collins et al., 1989; Tawfik et al., 2018). Importantly, the present section does *not* introduce fading as a novel invention, rather it is an outcome of the broader TRR 318 programme, and in particular in the project-A05 line of work where fading has already been treated as a guiding idea for adaptive explanation and tutoring design (Rohlfing et al., 2026). The aim here is to *operationalize* that shared idea for the present thesis within contrastive framework and to make it precise enough to support modeling decisions.

Accordingly, this scaffolding section goes beyond what the accompanying papers cover: it proposes a compact framework that links (i) an estimate of the learner’s evolving state (monitoring) to (ii) a policy for how much information is made explicit over time (fading). The purpose is to clarify what it would mean, computationally, for contrastive guidance to begin as strongly supportive and then gradually withdraw as the interaction succeeds.

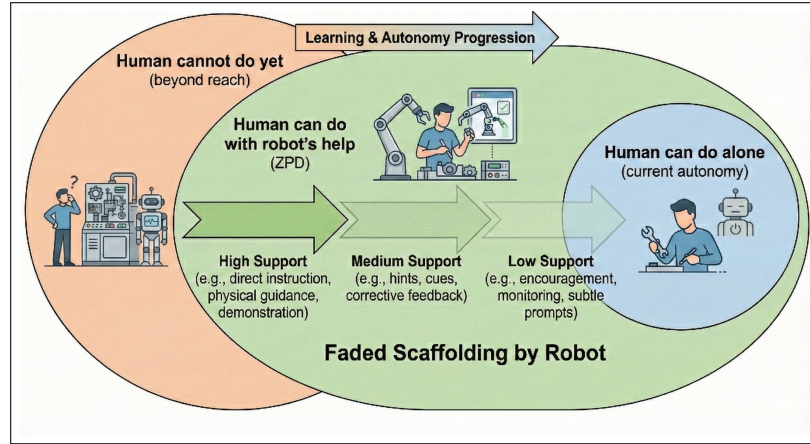


Figure 4.2: Faded scaffolding adapted to human-robot interaction. The diagram illustrates the application of Zone of Proximal Development (ZPD) to an HRI context. The robot acts as the More Knowledgeable Other, providing high-level support (scaffolding) when the learner is within the ZPD. As the learner internalizes the task and moves toward the center (Independent Ability), the robot's guidance is systematically reduced ("faded") to promote autonomy.

The next section therefore elaborates faded scaffolding and sketches how it can be implemented in the current work with the potential to be used in modeling. Following that, the previous ContACT-R model is extended into an interactive model of scaffolding (InterACT-R), in which context is dynamic, performance feeds back into the choice of guidance, and guidance is progressively reduced as learning advances.

4.3.1 Faded scaffolding

One of the core ideas of scaffolding is to adjust to the learner's ability and provide tailored support (Wood et al., 1976). Whereas the question of tailored support depends on how the linguistic instruction and its information presentation are controlled, the learner's ability is a very dynamic concept which is continuously changing (Vygotsky, 1978). For example, in the beginning, a learner can be perceived to be without a clear understanding of the task, but after some time she may develop a relatively better understanding. Here, one shall ask: how does a tutor react to such changes and, in return, adjust their strategy to align with the trajectory of learning? When we raise this question, it seems that we are exposed to two problems to be solved: a) Where lies the distinction between instructive and self-driven phases? Often we see that after some effort from both ends, learning evolves on its own; the question then is where this boundary lies and how the dynamics evolve. And, b) how can one characterize the relationship between learning and faded scaffolding? Here, the core problem essentially is to find a function which relates the rate of learning to the rate of withholding information along the learning trajectory. In this section, I will try to

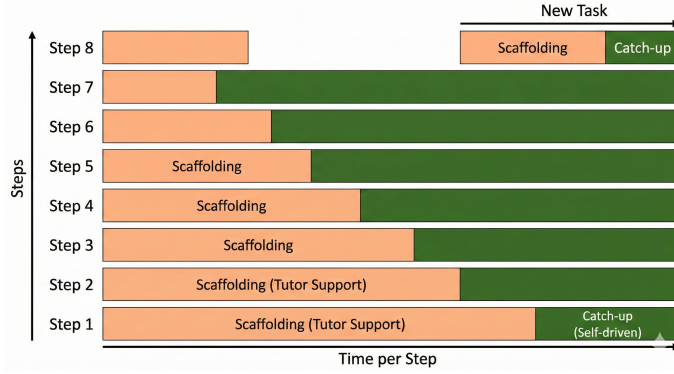


Figure 4.3: Figure 4.3: Distribution of tutor support versus learner autonomy, illustrating the fading effect as the step in the task increases. The fading resets as the new task is introduced.

address these questions one by one in order to find some formalization methods which have the potential to be implemented in HRI settings.

Addressing the first problem, it has been found that an efficient tutor, in addition to providing tailored instruction on how to present specific information, also regulates the amount of information needed for the learner to just “catch on” to the task (VanLehn, 2011). Such a state, where the learner is left to infer or execute the rest of the steps to solve the task, has been found to foster more learning as opposed to when the tutor leads the learner all the way to the solution (Chi, 2009). This concept, often referred to as the *assistance dilemma*, suggests that withholding information can be as crucial as providing it (Koedinger & Aleven, 2007, p. 242). The idea is powerful and has implications for learning, yet remains largely unutilized in current tutoring systems; hence, it is futile without being concretely implemented from this perspective. The idea has crucial implications for tutoring robots, for instance, which need to decide when and how much instruction should be provided for optimal learning—perhaps even more important than how the information is constructed (such as contrastive or not). Imagine a scenario where a learner is performing the task efficiently and comes to understand the whole task by themselves after certain time steps, triggering the “catching on” phase. Essentially, the two phases, namely, *scaffolding* and *catching on*, can be distributed along a timeline yet remain difficult to disentangle.

The question we shall ask then is: After how long within the task, and after how much instruction, is a particular individual able to catch on to the task so as to solve it without any subsequent help? Essentially, to test such questions, the experimental setup should at least fulfill some requirements. Firstly, a system should allow instructions to be implemented iteratively in steps such that, along the line, the steps are measurable points after which the learner is able to attain the “catch up” state. Secondly, the system must have the ability to be controlled for holding back predefined information (deliberate omission) based on some importance structure defined in terms of task

model (an obvious example being Path (“where”) components being more important than Manner (“how”) components for minimal task completion (Talmy, 2000), though the ideas should be generalizable to other tasks as well).

If an experimental setup provides these two abilities, then one can show that the dynamical tradeoff, and therein the attainment of learning, can be measured based on individual capabilities and the distribution of self-driven versus supportive learning phases. Though very general and not profiled for a specific learner, an example visualization of task attainment states can be realized as provided in Figure 4.3. During the first few steps, it can be assumed that the instructional phase lasts longer than the “catching on” phase, and as the task unfolds with time, the distribution should shift so that the self-driven phase lasts longer than the minimal instructional support. Whereas if the task remains similar on some dimension, the overall behavior profile is expected to exhibit a consistent and predictable shift; however, if a change in some task dimension is introduced, or the task changes entirely, the profiles should reset themselves since new instruction will be needed for enabling (Figure 4.3).

The second problem targets complementary aspect of scaffolding which concerns characterizing the relationship between fading (and the amount thereof) and learning (Collins et al., 1989). Although the solution appears to be simple due to its linear framing in language, such as “fading should be directly proportional to the rate of learning” (i.e., the quicker one learns, the quicker fading should follow), the solution is not simple, and the co-evolution of such learning and the connected scaffolding style is often convoluted. To start with, the problem can be decomposed into two dimensions: on one axis, task understanding, and on the other, fading, as visualized in Figure 4.4. Now the problem here is to find, at least as expressed in words, a function that characterizes both dimensions, crucially, for a particular individual and for specific task.

Before moving forward with the relations on two dimensions, it makes sense to recapitulate the meaning of these dimensions. Understanding is related to “task awareness,” which captures not only the current or instantaneous performance for the task but also ensures proper “enabling” such that the participant feels empowered and more confident for future performance (Bandura, 1977; Buschmeier et al., 2025). If one is looking for a more concrete and measurable term, then understanding can be defined (from our previous example and visualization in Figure 4.4) as the steps required by the learner to cover more than some prespecified threshold, let’s say an 80% “catch up” region. This definition fits well with enabling since we observe here, in the real sense, the learner internalizing the task with minimum support from the tutor; when the tutor fades the scaffolding, the learner boosts confidence through self-performance. Following this definition, fading should characterize the amount of information that is held back by the tutor; essentially, the “amount” should be decided by the relevance structure within the instruction as it appears (for example, Path information hierarchically being more important for the task than Manner).

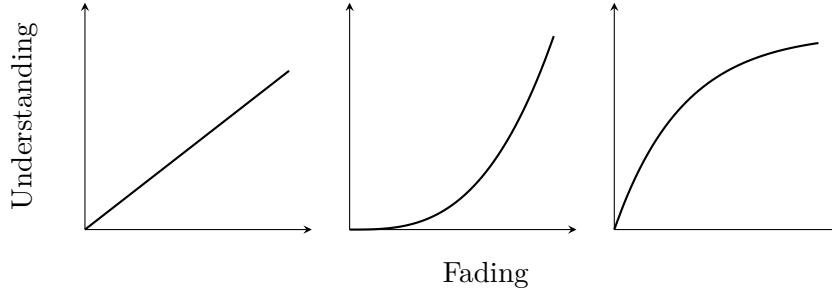


Figure 4.4: Illustrative relations between degree of scaffolding fading and learner understanding.

It seems that, after taking the rate of attaining “understanding” into account and the rate at which fading is introduced by the tutor, three possible scenarios should emerge as visualized in Figure 4.4. The first obvious scenario one can imagine is that fading is linear to understanding; in such a case, the amount of information decreases linearly as understanding evolves. Two other relations one could imagine (which look more realistic) are where the fading and rate of understanding are non-linearly related. One other possibility that we have not yet thought about but is possible involves dynamical systems, such as a Lorenz attractor, where small changes to the initial point lead to entirely different output states (Lorenz, 1963). Nonetheless, the three scenarios, together with the HRI setup and controlled parameters, provide us a very promising pathway to test not only the above-mentioned scenarios but also to derive the function fitting individuals by profiling their learning for an emerging task and context. The next section builds on some of these ideas to model contrast in ContACT-R model.

4.3.2 An interactive model of scaffolding

Recall that in our previous ContACT-R model (Chapter 2), our aim was to derive a simple rule for estimating relational prominence $R_d(C)$ using two conceptually grounded representations—namely *Ground* and *Figure* (Chapter 1), which was based on their conceptual distance and scaled by the alternatives, if provided in the scene. Importantly, this type of model was able to capture relational prominence $R_d(C)$ in a static context, allowing for a selection of implicit or explicit contrast as guidance. Whereas a static context provides a favorable ground for testing our model without adding unnecessary chaotic behaviors, in reality, the context itself has a “dynamic” component, as we have seen in the previous chapter (Dourish, 2004). We have also established that although processes related to language can be studied effectively in traditional psycholinguistics experiments where the context is held static (e.g., in Visual World Paradigms), the reality is more complex. Traditional approaches often fail to capture what interactionist approaches are able to holistically appreciate in dynamic contexts (Goodwin, 2000). To this end, we also highlighted that scaffolding is an integral process

within an interaction; interaction does not imply raw turn-taking without intention, but rather a process where partners monitor each other’s understanding and use interaction to scaffold each other (Rohlfing et al., 2021). Meaning, scaffolding is a central part of any interaction, not merely an outcome of turn-taking behavior, but a requirement for meaningful coordination to exist (Wood et al., 1976).

Nevertheless, the first step to capture the dynamic context is to allow the model to flexibly adapt as the context unfolds over time. Following the previous chapter on the contrastive ACT-R model (ContACT-R), we add a time dimension such that the model will now estimate the relational prominence $R_d(C_t)$ at each time t as the context unfolds.¹ Here, d suggests each relevant dimension (situative, functional, locative, eventive, and social), which can change during the task. The exact resolution window of t , i.e., whether it represents each step, each task, or each instruction, is not defined yet; however, once sufficient data is available, the estimation of t should not pose a difficulty for further modeling work. Hence, the same $R_d(C)$ from the ContACT-R model can be extended to the InterACT-R model as $R_d(C_t)$, but time-indexed, allowing it to change slowly because the context and parameters adapt over time, as given in:

$$R_d(C_t) = \Delta_d(C_t) A_d(C_t) \quad (4.1)$$

According to this, at each time step, the tutor observes a task context C_t and computes relational prominence $R_d(C_t)$, then decides whether to produce an explicit, contrastive guidance (e.g., negation) on dimension d or to stay more implicit. The solution seems simple thus far; however, up to this point, I have treated scaffolding mostly at the level of conceptual ideas rather than implementation through learning and interaction. Whereas contrastive guidance is a way of packaging information, we now have the notion of *faded scaffolding* as a way of regulating how much of that information is actually given to the learner at different points in time (Collins et al., 1989). Hence, we need to connect this model to learning. For learning, to start with, we can follow the Rescorla-Wagner learning equation, which is based on updating future estimates based on current outcomes (Rescorla & Wagner, 1972). In this system, it tracks a very simple performance signal: Let $e_t \in [0, 1]$ denote an error measure at time t ($e_t = 0$ means perfect performance, $e_t = 1$ a complete failure). From this raw signal, we construct a smoothed notion of performance P_t at time t .

$$P_t = P_{t-1} (1 - \gamma) + \gamma (1 - e_t), \quad (4.2)$$

Here, $\gamma \in (0, 1]$ is a smoothing parameter (or learning rate). Intuitively, P_t can be read as “how well the learner seems to be doing at this point,” with low values for sustained

¹In the static case, relational prominence was defined as a time-independent term $R_d(C)$, calculated as a product of relational distance and alternative availability; here I simply assume that this has already been computed by the model and can change over time with the task.

failure and high values for sustained success. Formally, since P_t follows a delta rule akin to Rescorla–Wagner, the current estimate is corrected towards the most recent outcome $(1 - e_t)$ with learning rate γ (Sutton & Barto, 1998). With a high learning rate γ , a single correct outcome will have a strong impact on P_t , such that the estimate is rapidly pulled towards 1; thus, the model quickly interprets the learner as having high understanding. In terms of the faded scaffolding policy, this leads to a fast reduction of support after only a few successful moves, making the tutor highly reactive but also more vulnerable to noisy performances or success by chance. Hence, the problem boils down to effectively estimating γ for each individual, rather than keeping it constant across all learners. Since γ plays the role of a learning-rate or memory parameter, we can allow a participant-specific value γ_i for each participant i , and estimate it from each learner’s behavior (e.g., using a Bayesian hierarchical model) (Gelman et al., 2013). Learners with larger γ_i will be modeled as updating their performance estimate P_t^i more quickly based on recent outcomes, whereas learners with smaller γ_i integrate evidence more slowly over many attempts or trials. This fits naturally with the idea of profiling individual learning trajectories in our HRI setting as designed in Singh and Rohlfing (2025b). For the present purposes, we treat this smoothed performance directly as the model’s estimate of *understanding* and write:

$$U_t = P_t, \tag{4.3}$$

Following this, $U_t \in [0, 1]$ corresponds to the “understanding” axis in Figure 4.4. At this point, nothing in what follows depends on the exact definition; if a better or richer measure of understanding becomes available, it can simply replace P_t .

Formalizing the Fading and Understanding relationship. In this section, I will attempt to provide a solution to the question raised in Figure 4.4. The core problem here, as noted before, is to find an optimal function relating *Understanding* and *Fading*. As we have already derived some use cases for how understanding U_t and fading F_t might be related across two dimensions, and since U_t is now already known or estimated, we must derive the fading F_t as the next step.

Going through a simple solution, we start with a situation where scaffolding is either provided (1) or not (0). Let us denote this Scaffolding by S_t , where $S_t \in [0, 1]$. Hence, the second internal state, *fading level* F_t , can be defined by:

$$F_t = 1 - S_t. \tag{4.4}$$

Here, fading is essentially the absence of scaffolding. Although previous sections conceptualized fading as a graded process with information being held out depending upon the hierarchical relevance of information to the task (VanLehn, 2011), here, for simplic-

ity, we keep a binary choice of either providing or omitting the entire information. Since the equation $F_t = 1 - S_t$ encodes this relation, we will be able to add more scaffolding levels S_t (e.g., each information element containing a continuous range between 0 and 1) to introduce more fine-grained adjustments to F_t later. Nevertheless, at this point, $S_t = 1$ means full scaffolding (nothing has been faded yet) and $S_t = 0$ means that the tutor has withdrawn support as far as possible.

In terms of Figure 4.4, F_t corresponds to the “Fading” axis: the more we move to the right, the higher F_t becomes and the less explicit help is provided. In this graph (Figure 4.4), we also have highlighted a central problem, i.e., identifying a function mapping the fading state F_t to the estimated understanding U_t . This is where the different “profiles” of faded scaffolding, as we have so far imagined, help us make the concept concrete. In very simple terms, the model can assume a policy function $g(\cdot)$ characterizing the relationship between Fading F_t and Understanding U_t (Murray, 1999), such that:

$$S_t = g(U_t), \quad F_t = 1 - g(U_t). \quad (4.5)$$

We already know that the three outcomes shown in Figure 4.4 are a result of different choices of $g(\cdot)$. The function $g(\cdot)$ essentially yields different characteristic shapes for three illustrative cases. Theoretically, one can derive these shapes based on our assumptions, where for each case, we can see these shapes emerging when the following requirements are fulfilled:

Linear Fading (Figure 4.4, left): Here, the Fading F_t increases in direct proportion to Understanding where every step gain in understanding is accompanied by an equal amount of fading representing a constant withdrawal rate relative to understanding.

Accelerating Fading (slow at the beginning, fast later; Figure 4.4, middle): In this case, the tutor keeps support relatively high while the learner is still struggling, and then fades more aggressively once understanding picks up. This aligns with conservative tutoring strategies that prioritize error avoidance early on (J. R. Anderson et al., 1995).

Diminishing-returns Fading (fast at the beginning, then plateau; Figure 4.4, right): This pattern reflects a situation where the tutor takes advantage of early improvements to reduce support quickly, and then makes only small further adjustments as the learner moves toward perfection. This strategy is often observed in scenarios where the “Assistance Dilemma” presses for early withdrawal to encourage active learning (Koedinger & Aleven, 2007).

In practice, we cannot claim that one of these three forms is the correct one. Rather, they provide a compact use-cases of expressing different intuitions about how fading and understanding might co-evolve for different individuals and tasks (Rafferty et al.,

2016). Given an HRI setup that can record performance over time and manipulate how much information is withheld, one could try to fit such a function $g(\cdot)$ to actual learner trajectories. Since the problem is now defined, one could run a real study on this and estimate the optimal $g(\cdot)$ empirically.

Bringing relational prominence and understanding together Recall that the original ContACT-R model used relational prominence $R_d(C)$ together with a fixed threshold θ_d to decide whether an explicit contrast on dimension d should be produced.² One could apply the same rule here, provided the task were not dynamic, as in the ContACT-R model. In contrast to the scenario for the previous ContACT-R model, here the participant is also acting on the task. Thus, in addition to relational prominence $R_d(C_t)$, we must consider task performance, given by learning P_t (Eq. 4.2), or a more generalized form of understanding U_t (Eq. 4.3) (Corbett & Anderson, 1994). Hence, we need a function that considers both relational prominence $R_d(C)$ and the understanding parameter U_t to form sufficient evidence for using contrastive guidance. Let us call this “evidence” term $E_{d,t}$, which depends on both relational prominence $R_d(C_t)$ and U_t , evolving with time t during the interaction (Ratcliff, 1978). In the interactive setting, this decision can now be made sensitive both to the learner’s current understanding U_t accumulated so far and to the relational prominence $R_d(C_t)$. Here, we can define the continuous *Evidence* term as:

$$E_{d,t} = \kappa_d R_d(C_t) + \eta_d(1 - U_t), \quad (4.6)$$

In this equation, $R_d(C_t)$ indicates how contrastive the current context (C_t) is along dimension d . We also introduce a weight parameter κ_d for relational prominence, which essentially weights the general importance of $R_d(C_t)$. A high κ_d implies that even if the prominence between *Figure* and *Ground* is low, the model will push for guidance, since κ_d is high. On the other hand, when κ_d is small, even a high $R_d(C_t)$ doesn’t matter much. Up to this point, the equation differs little from what we utilized in the previous ContACT-R model, where only relational prominence mattered for the decision.

However, here we have the additional term $\eta_d(1 - U_t)$, which relates to performance. In this equation, U_t represents understanding by capturing the learning trajectory P_t over time t (Eq. 4.3). Crucially, rather than measuring understanding as all-or-none (0, 1), P_t is a smooth estimate that changes gradually. If the learner is doing well overall (high P_t), the term $\eta_d(1 - U_t)$ is small, resulting in low evidence for guidance. Conversely, if the learner is struggling over several trials (low P_t), $\eta_d(1 - U_t)$ becomes large and steadily pushes $E_{d,t}$ up, making guidance more necessary (VanLehn, 2011). Similar

²In the previous ContACT-R model, we had first chosen a hard decision rule, where if $R_d(C) > \theta_d$, contrastive guidance was selected, or a probabilistic version in which the probability of selecting negation increases monotonically with $R_d(C)$, captured through a logistic function.

to κ_d , η_d is a weight parameter determining the importance of performance for the evidence $E_{d,t}$. If one wants to be very sensitive to even subtle changes in performance, one can set it high; otherwise, to remain inert to small changes, the value can be kept low.

Deriving the threshold for guidance Since we now have the evidence parameter $E_{d,t}$, we need a decision rule for contrastive guidance to be triggered depending upon this evidence. In the ContACT-R model, evidence was solely based on relational prominence, and guidance was triggered as soon as this evidence exceeded a threshold value θ_d , which was either fixed or derived experimentally on a probabilistic scale. Here, however, the value of $\theta_d(t)$ is time-dependent and should also account for how much Fading F_t has been achieved. A high Fading F_t suggests that, to trigger guidance, the evidence should be stronger compared to when Fading is low.

Let us assume that our initial threshold for guidance is $\theta_d^{(0)}$. Triggering guidance initially would be an easier task; however, later in the interaction when Fading is higher, the same evidence should no longer trigger guidance, and the learner must significantly “struggle more” to receive it. Hence, let us call this new threshold after Fading $\theta_d(t)$. Crucially, this new threshold $\theta_d(t)$ can simply be conceptualized as how much the threshold shifts as Fading increases, given by:

$$\theta_d(t) = \theta_d^{(0)} + \omega_d F_t. \quad (4.7)$$

Here, $\theta_d^{(0)}$ is the base threshold in the early phase of learning, and $\omega_d \geq 0$ specifies how strongly the threshold on dimension d rises as scaffolding is faded. Had we used a hard decision rule, we could have compared this $\theta_d(t)$ with evidence $E_{d,t}$ to trigger the guidance decision simply by:

$$\text{Guide on dimension } d \text{ at time } t \text{ if } E_{d,t} > \theta_d(t). \quad (4.8)$$

Here, instead of comparing $E_{d,t}$ to a fixed threshold θ_d , we let the threshold itself depend on the fading level $\omega_d F_t$. In simple terms, $E_{d,t}$ asks: given the current context and performance, how strong is the case for guidance? Meanwhile, $\theta_d(t)$ asks: given how far we’ve faded (F_t), how high is the bar for acting in this case? According to this, as F_t (fading level) increases, $\theta_d(t)$ rises, so the same evidence $E_{d,t}$ that would have triggered guidance early on will no longer be sufficient later during learning

In an intuitive sense, this means that in the beginning (when F_t is small and $\theta_d(t)$ is close to $\theta_d^{(0)}$), the system is relatively willing to use contrastive instructions whenever prominence and low understanding jointly suggest they are needed. As understanding increases and scaffolding fades, F_t grows, $\theta_d(t)$ increases, and explicit contrast now requires much stronger evidence to be triggered. The learner is pushed into a more

self-driven “catch-up” phase, where only particularly long pauses or divergence will lead to overt guidance.

While we have discussed a hard decision rule so far, where at time t , if $E_{d,t} > \theta_d(t)$, guidance is triggered, we can instead define the probability of choosing an explicit, contrastive guidance move on d at time t by a logistic choice rule, similar to the ContACT-R model. This is based on evidence $E_{d,t}$ and threshold $\theta_d(t)$, with the core input $(E_{d,t} - \theta_d(t))$ to the sigmoid function:

$$\Pr(\text{contrast on } d \mid C_t, U_t) = \sigma(E_{d,t} - \theta_d(t)) = \frac{1}{1 + \exp(-E_{d,t} + \theta_d(t))}. \quad (4.9)$$

By doing this, we are essentially transforming the hard yes-no decision rule into a graded likelihood of providing guidance. To summarize, the interactive ACT-R model (InterACT-R) of scaffolding proposed here extends the earlier ContACT-R framework in three ways.

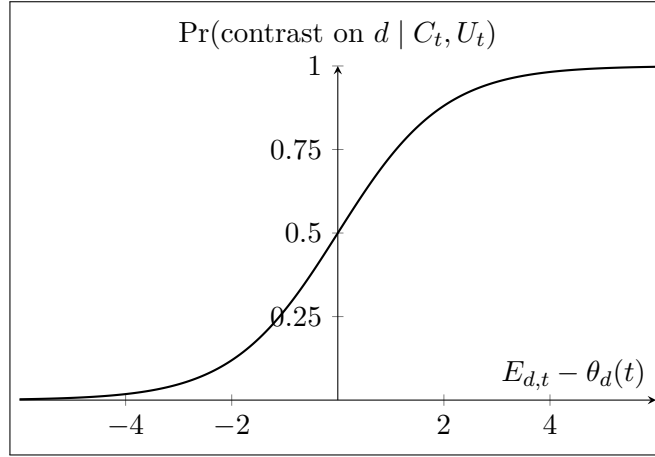


Figure 4.5: Logistic mapping from evidence minus threshold to the probability of contrast on dimension d .

First, it adds a time-varying estimate of understanding U_t based on observed performance. Second, it introduces a fading state (S_t, F_t) that links this understanding to the amount of scaffolding the tutor provides, with different functional forms corresponding to different fading–learning profiles as in Figure 4.4. Third, it lets this fading state modulate the decision to give contrastive guidance by changing the threshold for explicit assistance over time. Although these ideas are yet to be implemented in a real setting, the formalization provides a concrete pathway to test the core concepts behind scaffolding, especially in contrastive contexts. The aim here was to make these additions to bring the model closer to the informal picture developed in the previous subsection: a tutor that not only knows *what* to provide (contrasts) but also dynamically decides *when* to withdraw and let the learner “catch on” to the task by themselves.

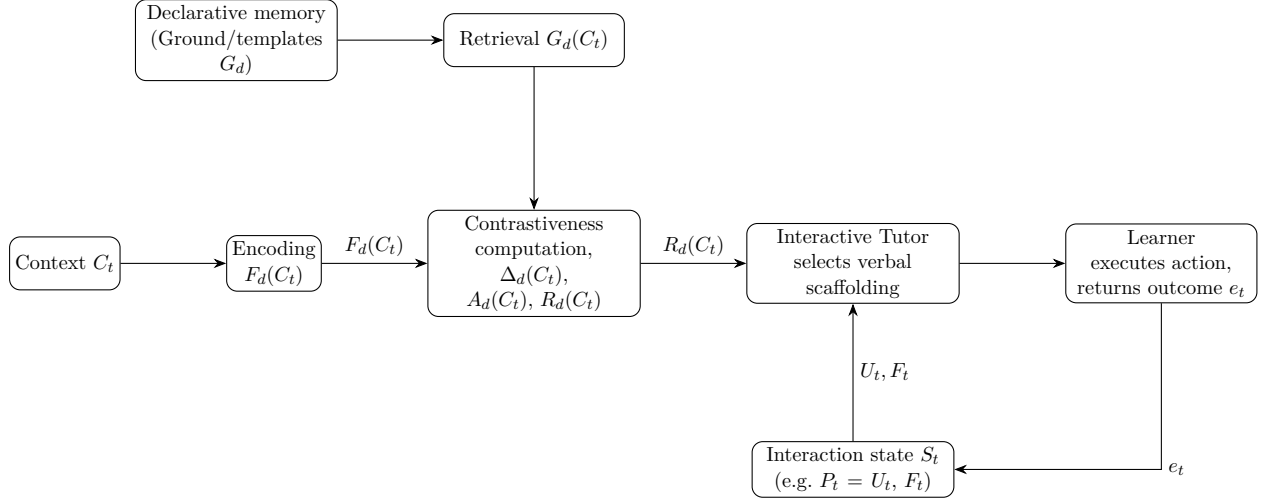


Figure 4.6: InterACT-R as an interactive extension of the ContACT-R model. The left-to-right backbone (Context $\rightarrow$ Encoding $\rightarrow$ Contrastiveness) is preserved; an interaction state S_t and task feedback loop modulate the guidance policy over time.

Interim summary This interactive extension of ContACT-R remains deliberate hard-coded, but it already changes the interpretation of the previous (rigid) model in important ways. First, it makes room for *learning on both sides*. The learner’s behavior is still explained in terms of perception, retrieval, and action selection, but now the tutor’s behavior also evolves where guidance is no longer a fixed script and it becomes a learned policy that adapts to the learner’s performance and to the kinds of contexts that are actually encountered (Thomaz & Breazeal, 2006).

Second, it ties scaffolding back to the core idea of contrastive contextualism. The same relational prominence values $R_d(C_t)$ that were previously used to define in ContACT-R when negation will be appropriate in a one-shot setting now also serve as state variables that guide the tutor’s choices in an ongoing interaction within the InterACT-R framework. In this sense, the previous ContACT-R model (Figure 2.9) can be extended into an InterACT-R model by including a notion of feedback loop from the learner to the tutor as shown in Figure 4.6.

Third, it connects language and task performance in a principled manner. Performance indicators such as e_t and P_t formally link the success or difficulty of actions to the choice of linguistic strategy. Whereas repeated errors increase the estimated need for scaffolding, successful independent actions gradually reduce the scaffolding needs. This makes it possible, at least in principle, to simulate how a contrastive scaffolding policy would evolve over time for learners with different initial knowledge or different patterns of mistakes (VanLehn, 2011).

4.3.3 Discussion

The aim of this chapter was to bring together the idea that the previous chapters had introduced. Here, the aim was to move from contrastive guidance as an isolated utterance in a static context to contrastive guidance as part of an *interactive* scaffolding process. Doing so required a divergence from studying the dominant idea of turn-taking to focusing purely on scaffolding. This divergence was important because studies in interaction often heavily focus on systematizing turn-taking processes by looking into turn timings and places where turns are more likely to happen (Levinson, 2016; Skantze, 2021). In this process, we seldom ask why the turns exist in the first place and what their purpose is. It turns out that mostly, turns are not for the sake of keeping the conversation going; rather, they are an act of scaffolding the partner, where the aim is to enable a partner for the task in hand (Buschmeier et al., 2025; Rohlfing et al., 2021).

This particular notion is also important due to its implications for how we study language processes themselves, as highlighted in the previous chapter where studying context also provided two distinct notions. Not surprisingly, these two traditions are reflected in language research. On the one hand, there is the traditionalist or cognitivist approach, where language processing is treated as an internal computational problem with well-defined inputs (Fodor, 1983). On the other hand, there is the interactionist approach, where language is seen as a joint activity in which context is developed as people act, speak, and repair each other (Goodwin, 2000). The central claim of this chapter has been that neither side, taken alone, is sufficient if we want to understand how contrastive guidance can be used as scaffolding in real tasks, and that Human–Robot Interaction (HRI) offers a useful meeting ground for bringing them together (Singh et al., 2025).

We began the chapter by unpacking these two sides more explicitly. The traditionalist stance formed the ground for most psycholinguistic work and the original ContACT-R model by fixing context as a static Ground within each trial and focusing on the internal dynamics of a single mind, providing us with control and formalization (J. R. Anderson et al., 2004). This was a promising start, which later turned out to be incomplete on theoretical and practical grounds. Here, we were convinced to say contextual factors are already represented, and nothing was developing as the interaction unfolded.

Following this, the interactionist stance started from a different intuition: Context is not merely “what is available”, but also “what is being developed as part of interaction” between partners (Schegloff, 1992). We also proposed HRI as a methodology to study language and visual processes and treated conversational repairs, gaze, and action performance as proxies to study visuo-linguistic processes online (Singh & Rohlfing, 2025b). Unlike classic Visual World Paradigms, which have been an established field to study the bidirectional relation between language and vision, and where participants are confined to looking at static contexts (Tanenhaus et al., 1995), HRI provided

an opportunity for participants to act, change the context, and interact with a partner through repairs and receive scaffolding. Essentially, an *interactional context* developed as a process provided both static and dynamic elements where language and visual processes could be studied in a controlled experimental setup (Singh & Rohlfing, 2025b).

Whereas for linguistic processes one could show that the use of controlled omission to elicit repairs (later known as the Fading strategy) can be used as a window into real-time language-visual correspondence, on the non-linguistic side, gaze was treated not only as an indicator of referential activation but also as a trace of how the evolving context was proactively used as an external resource (Admoni & Scassellati, 2017). In this way, visual and language behavior was made visible within interaction without compromising on the task continuity by selectively targeting static and dynamic aspects of context in real-time.

As a synthesis of the preceding strands, the final section formalizes and extends the old static contrastive (ConACT-R) model to an interactivity- and scaffolding-capable InterACT-R model. Although the basic ideas of the original model were retained, such as context C_t represented as a set of dimension-specific Grounds $G_d(C_t)$, Figures $F_d(C_t)$, and relational prominence values $R_d(C_t)$ that quantify how far the Figure lies from Ground, a notion of interactional time t , performance measures e_t and P_t that summarize how well the learner is doing, and a set of guidance strategies whose utilities are updated as the interaction unfolds were included to make the model interactive (VanLehn, 2011). Thus, in the extended model, contrastive guidance remains no longer a fixed utterance, but rather a strategy that adapts to the learner’s performance and fades scaffolding.

On this note, this interactive InterACT-R sketch is only a first attempt where many complexities, especially concerning scaffolding, remain unaddressed. For example, the learner’s own internal learning dynamics are treated in a highly simplified way through performance at a specific time. The robot’s side of the interaction is also idealized for the best-case scenario; in practice (as seen in anecdotes from my experiment), real systems have limitations in speech recognition, timing, and embodiment that can also shape how scaffolding is experienced (Thomaz & Breazeal, 2006). However, I do not see these omissions as reasons to abandon the modeling effort, but as a reminder that formal accounts of scaffolding will have to grow in tandem with richer empirical work if they are to remain psychologically plausible.

Nevertheless, this chapter has argued that any serious theory of contrastive guidance and scaffolding must accommodate two kinds of constraints at once. On the one hand, constraints imposed by the cognitive processes, i.e., memory retrieval, attention, and decision-making putting limits on how contrastive cues can be processed and used. On the other hand, there are constraints imposed by interaction: context is co-constructed, misunderstandings are collective events that need to be managed, and guidance has to be timed and faded in ways that fit an ongoing activity (Rohlfing et al., 2026).

Chapter 5

Conclusion and outlook

From Negation to Contrast and Context, the title already tells a story of how this thesis has developed over time. I started as a psycholinguist interested in negation as a rather local linguistic phenomenon, mostly following established questions such as why and through what mechanisms negations are more difficult to process than their affirmative counterparts (H. H. Clark & Chase, 1972; Kaup et al., 2006). Over the course of the project, this narrow starting point expanded into a broader investigation of contrast as a relational principle, and eventually into a view of contrast as one way of understanding context itself. In this sense, the thesis can be read as a gradual movement from a single linguistic device toward a more general account of how relations between Figure and Ground are structured in perception, language, and interaction (Talmy, 2000).

The first step in this trajectory can be understood as conceptual. In the introduction, I proposed contrastive contextualism as a way of capturing the intuition that our perception is almost always relational and that many so-called contextual effects can be understood as graded differences in relational prominence between a Figure and its experiential Ground. Rather than treating contrast as something exclusively encountered in either linguistic or visual processes, I was convinced to argue that contrast might be a pervasive feature of everyday situations—which we call contexts—that becomes visible when the Figure–Ground relation departs from what experiential traces would prepare us to expect. This motivated me to articulate specific dimensions along which I could imagine such contrasts emerging; although not exhaustive, I proposed situative, functional, locative, eventive, and social dimensions. Together, these dimensions provided a first foundation for talking about contrast as a form of context, and for locating different phenomena on a common contrastiveness scale.

Provided this conceptual foundation, I then turned to the question of how language might interact with these relational dimensions, and whether and how linguistic expressions can act as mechanisms for structuring relational perception. I argued that everyday linguistic devices, in particular the packaging of motion events and the use of negation, can shape ongoing event conceptualization by spotlighting specific relational alternatives and augmenting the space of available contrasts (Verhagen, 2005). Supplemented by empirical studies on motion events and on functional manners in Hu-

man–Robot Interaction, I attempted to support these theories further at the end. One finding stood out particularly: when verbal guidance brings past and present sub-events into a single frame of interpretation—such as through contrastive instructions involving negation—observers are more likely to encode the event as a structured whole (Singh & Rohlfing, 2025a). This finding essentially broadened my approach to see negation beyond sentences, framing the initial linguistic question about negation toward a more open, general question on how negation can serve as contrastive guidance for general cognition and learning.

With context serving as the foundation for all these developments, the next step was to bring it within the scope of discussion. Building on past literature, context was distinguished between treating it as “what is available” and “what is developed” (Dourish, 2004). Unsurprisingly, I observed these two stances running in parallel through the study of language as well, which later also formed the background for a methodological paper on studying online language and vision processes in HRI (Singh & Rohlfing, 2025b). On the contextual distinction, there is a tradition that treats context as static and focuses on computation. This is in contrast to an interactionist tradition that sees context as something co-constructed within interaction (Goodwin, 2000). The present thesis did not attempt to adjudicate between these positions definitively. Rather, the aim was to hold them together by showing that many of the phenomena of interest, including contrastive guidance, require both a representational perspective—where we can talk about stored Grounds and Figures—and an interactional perspective, where we can observe how those Grounds are reshaped through action, repair, and scaffolding.

As a result of this synthesis, modeling work became possible. Once contrastiveness had been described in phenomenological terms and explored experimentally, the question arose whether these ideas could be made explicit in a computational framework. The first answer was the ContACT-R model (or Contrastiveness ACT-R), which treated context as a set of dimension-specific Figures and Grounds and defined relational prominence as a graded function of their distance and the availability of alternatives. This static model was deliberately rigid, as it did not capture interactivity, but it was sufficient as a starting point to show that the informal notion of Figure–Ground distance could be implemented in a form compatible with existing cognitive architectures (J. R. Anderson et al., 2004). In doing so, ContACT-R served as a bridge between the conceptual analysis of Chapter 1 and the experimental results on linguistic guidance, offering a first technical handle on when explicit contrast, including negation, is warranted. However, as the thesis progressed, the limitations of a purely static treatment of context became increasingly clear. The Human–Robot Interaction studies showed that many of the phenomena I was interested in—such as scaffolding, Fading, and Monitoring—only fully reveal themselves when context is allowed to change through the learner’s and tutor’s actions (Singh et al., 2025). This motivated an extension of the modeling framework toward an interactive version, InterACT-R, in which contrastive

guidance is no longer a one-shot decision but a part of interaction that adapts over time. In this extended model, relational prominence values continue to play a central role, but they are now indexed to interactional time, coupled to estimates of the learner’s understanding, and linked to fading functions that regulate how much help should be provided (Collins et al., 1989). The model remains a sketch, but it illustrates how the basic ideas of contrastive contextualism can be carried into a setting where learning histories, repairs, and scaffolding moves begin to matter more.

Throughout this development, the empirical studies that make up the individual papers of this thesis have been more than isolated demonstrations, providing concrete contributions to the sections in which the abstract ideas were realized. For instance, the study on motion cognition provided clear definitions of what counts as a “Ground” or “Figure” and how goal-bias interacts with mid-event structure (Singh & Rohlfing, 2025a). The touchscreen and manner learning task in Human–Robot Interaction offered insight into the functional dimension of contrastiveness, showing how routinized affordances can be used as experiential grounds for scaffolded alternatives (Groß et al., 2023). The HRI study on explanatory dialogue and fading scaffolding required thinking of contrastive guidance as part of an unfolding interaction, where repairs, gaze behavior, and action performance all become data for understanding how context is being used (Singh & Rohlfing, 2024). At the same time, it is important to highlight that the thesis is limited in several respects. For instance, the experiments supporting the ideas have been necessarily constrained, utilizing simple tasks and limited action spaces. The modeling leaves out many rich complexities of learning and interaction, such as how canonical Grounds are acquired over time, how multiple candidate Grounds compete, or how scaffolding occurs in the real world between a child and caregiver (Wood et al., 1976). Even at the conceptual level, contrastive contextualism has been developed mainly for a small set of dimensions and mostly in visually grounded tasks, leaving open how far the same framework can be extended to richer multimodal and social settings. These limitations do not undermine the central claims, but they do delimit the scope of the present thesis and point toward where further development is needed.

5.1 Limitations

Any attempt to bring together conceptual analysis, experiments, and modelling within a single framework is bound to remain incomplete, and this thesis is no exception. Hence, the first limitation can already be identified as being too focused; the thesis started from contrastive contextualism and moved through negation to contrastive verbal guidance. Although this focus has helped me to develop a coherent line of argument, akin to a story, it has also led to important aspects of context and interaction remaining unaddressed. For instance, contrastive contextualism has been worked out mainly for a subset of dimensions that could be tested only in well-controlled tasks, treating the

social dimension minimally due to a lack of results. In addition, the thesis has mostly operated within dyadic settings in which meta-communicative aspects such as power relations and roles are treated relatively sparsely, leaving more complex configurations for future work (Bourdieu, 1991; Foucault, 1980).

Methodologically and empirically, the studies that support the thesis were deliberately chosen to reflect contrast on specific dimensions, such as motion events and action manner manipulation; nevertheless, they rely on highly controlled scenarios which are yet to be verified on other specified dimensions. For instance, the motion events primarily depict locative contrast, and a more generalized notion of this concept is yet to be investigated. Similarly, the action performance in the Human–Robot Interaction study implemented constrained workspaces, limited sets of objects, and fixed instructional scripts, which compromised ecological validity to some extent (Jung & Hinds, 2016). Real-world interactions are inherently messy and multimodal, often defying the clean variables of laboratory settings.

The modelling work, both in ContACT-R and its InterACT-R extension, also highlights substantial limitations at this stage. The first limitation can be recognized at the representational level, where the Ground concept was treated as already given, abstracted away from how it is learned in the first place, which made the entity less dynamic (McClelland, 2009). In the interactive extension, this limitation was partly addressed; however, the partner and task models were still reduced to a few observable measures. This was also partly due to the model’s inability to implement rich multi-modal signals, which are an integral part of any real-world social interaction (Kendon, 2004).

Finally, the Human–Robot Interaction scenarios used to support the theoretical and empirical concepts are still based on systems that are not fully adaptable. Although the robot in the experiment was responsive and capable of providing contrastive guidance, many aspects of the interaction were pre-scripted, leading to systematic artifacts rather than natural interaction (Riek, 2012). Hence, the scaffolding and fading patterns being analyzed are inferred from behavior rather than being generated by an implemented version of InterACT-R running online. While this is appropriate for a first exploration of the space, it means that the thesis has yet to investigate how a real system could learn and adapt a contrastive scaffolding policy on its own. All these limitations should be seen as potential aspects for improvement, as they indicate where the current thesis draws its boundary and where further development is required.

5.2 Future directions

The limitations just outlined do not only mark what is missing; rather, they also help to shape the research direction where the thesis should extend in the future. Importantly, to develop a unified theory of contrastive guidance and its practical implications, it is

crucial to keep the three strands together in the future: conceptual robustness of the contrastiveness scale, empirical studies to evaluate those concepts, and refining models based on empirical outcomes.

A first avenue concerns broadening the empirical base of contrastive contextualism. The motion event and action demonstration paradigms can be enriched toward longer and more complex activities, where multiple subevents unfold over extended time (Sebanz et al., 2006). Tasks that involve planning, negotiation of roles, or coordination under uncertainty would make it possible to investigate how contrastive guidance operates when there is no single obvious Ground but a whole field of potential alternatives. Extending the studies to more naturalistic settings, including everyday interaction with complex tasks and collaborative problem solving, would test whether the patterns observed in controlled laboratory contexts generalize when scripts are less rigid. It will also be important to include more diverse participant groups, for instance, children, older adults, and speakers of different languages, in order to see how developmental histories and linguistic systems shape which Grounds become salient and how negation and other contrastive devices are used (Choi & Bowerman, 2006; Tomasello, 2008).

A second avenue concerns deepening the interactive perspective on context and scaffolding. The Human–Robot Interaction scenarios developed in this thesis have already opened a space in which context can be both controlled and developed, but this idea can still be taken further. The method for studying online language–vision interplay in HRI can be used as the basis for genuinely adaptive interactions, where the robot does not only deliver fixed contrastive instructions but adjusts its scaffolding and fading in response to the learner’s online behavior (Thomaz & Breazeal, 2006). Evaluating InterACT-R in such a setting would allow us to ask more precise questions about how different scaffolding strategies affect understanding and how fading relates to the emergence of autonomy. Beyond HRI, similar questions could be asked in human–human settings by combining conversation analytic techniques with eye tracking and detailed investigation of repairs (Bangerter & Clark, 2004; Goodwin, 2000).

The current version of ContACT-R and InterACT-R treat Grounds and Figures as given and do not include mechanisms for learning or updating them. Future work could integrate learning components that allow the architecture to acquire canonical Grounds from experience and to adjust similarity structures as new situations are encountered. This might involve combining ACT-R style declarative memory with Bayesian inference, so that the model can discover which contrasts are useful in a given task and how much explicitness is needed for a particular learner (Taatgen & Lee, 2013). Another potential area to explore further would be to integrate generative AI models like LLMs with the classical cognitive architecture such as InterACT-R, especially to learn Ground representations, an approach known as neuro-symbolic AI (Romero et al., 2024).

Moreover, on the interactive side, the partner model could be enriched to include not only estimates of task understanding but also qualitative measures, like satisfaction and

trust in the system (Muir & Moray, 1999). Implementing these ideas in robots with robust speech recognition through LLMs would make it possible to test how reliable the proposed concepts are and how they interact with the constraints of the model.

A fourth avenue points toward applications beyond the immediate experimental settings. The notion of contrastive scaffolding suggested in this thesis is closely related to practices in education and tutoring, where instructors routinely rely on contrasts, counterexamples, and negations (Aleven & Koedinger, 2006). The conceptual and modeling tools developed here could inform the design of educational technologies, from robotic tutors to intelligent tutoring systems, that make principled use of contrastive guidance and fading. Similarly, conversational agents and dialogical systems in everyday devices might benefit from moving away from a purely information-theoretic approach to interactions that explicitly help users to see relevant contrasts. In such applications, questions of transparency, trust, and user experience will become central, which in turn call for collaborations with emerging fields such as Explainable AI (XAI) (Gunning, 2017).

Across all these directions, the aim is not to abandon the core ideas of the thesis, but to put them under test in richer environments. Contrastive contextualism posits that context can be understood in terms of Figures and Grounds and that language can act as a mechanism for structuring their relations. The modeling work shows that this intuition can be made explicit, at least in simplified settings. The next step is to see how far these ideas can be extended, especially for complex social interactions, into systems that learn and adapt over time, and into applications where scaffolding is an integral part. In that way, the ideas presented in the thesis here can serve as a meeting point rather than an endpoint for a broader investigation of contrast, context, and scaffolding.

Appendix A

Relation to publications

A summary of how individual publications have contributed to the development of the present work. The ordering reflects the sequence in which their core ideas appear in the chapters.

Publications as they appear in the work:

- Singh, A. and Rohlfing, K.J. (2025), Contrastive Verbal Guidance: A Beneficial Context for Attention To Events and Their Memory? *Cognitive Science*, 49: e70096.

This journal article forms the empirical backbone of Chapter 2, where language is tested for its contrastive effects. Across two eye-tracking experiments on motion events, the study shows a complementary effect of visual and verbal contrast on action memory. It also demonstrates an inverse relation between attention to the goal and memory for mid-events, which can be mitigated by negation. Two methods of verbal-instruction synchrony are also introduced for the first time, showing how they modulate goal-attention. The systematic use of contrast along the locative dimension is developed in this paper, specifically where negation such as “not up” can bring the alternative path into focus. These findings are reinterpreted within the broader contrastive contextualism framework in Chapter 1 and are used to derive the design principles for contrastive guidance in Chapter 2. The study also motivates the initial notion of Figure–Ground relations and thereby relational prominence.

- Singh, A., & Rohlfing, K.J. (2023). Contrastiveness in the context of action demonstration: an eye-tracking study on its effects on action perception and action recall. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45, Sydney, Australia.

This conference paper reports an earlier phase of the motion-event work. Here, the focus was on assessing how negation can inhibit goal-bias in perception, thereby improving recall for the less salient parts of the event. The study shows that contrastive verbal guidance, especially when realized with an assertion–negation

combination, can direct attention toward the past sub-event and treat it as a Ground for interpreting the current action.

- Singh, A., & Rohlfing, K.J. (2024). Coupling of Task and Partner Model: Investigating the Intra-Individual Variability in Gaze during Human–Robot Explanatory Dialogue. *26th ACM International Conference on Multimodal Interaction (ACM ICMI 2024)*, Costa Rica.

This paper provides one of the central HRI studies upon which Chapter 3 and Chapter 4 build. It introduces the distinction between a relatively stable task model and a dynamic partner model within a controlled Human–Robot Interaction setup. Empirically, the study exploits the well-known asymmetry between Path and Manner: Path information is retrieved more quickly and forms a robust backbone for the task model, whereas Manner information is more variable and tightly coupled to the evolving interaction with the robot. By tracking gaze and repairs, the paper shows how these two layers together form what I call an interactional context in the current work, which is used in Chapter 3 to motivate the distinction between static and dynamic aspects of context.

- Singh, A., Wrede, B., Richter, B., Groß, A., & Rohlfing, K.J. (2025). Manners Matter: Action history guides attention and repair choices during interaction. *IEEE International Conference on Development and Learning (ICDL 2025)*, Prague. (Talk)

This publication elaborates on the idea that interactive contexts can be experimentally built through action histories and repairs. Within an HRI setup, the study manipulates whether the situational model on the table is simple or rich and examines how this history constrains both gaze and repair choices. In the thesis, these findings appear in Chapter 3, where history is treated as a resource that can be actively exploited. The paper also serves as a foundation and initial proof of concept for the claim that language–vision coupling in an online interaction can be studied in a controlled Human–Robot Interaction setup.

- Singh, A., & Rohlfing, K.J. (2025). Embedding Psycholinguistics: An Interactive Framework for Studying Language in Action. *6th Biannual Conference of the German Society for Cognitive Science (KogWis 2025)*, Ruhr-University Bochum, Germany.

This conceptual and methodological paper proposes an interactive framework for studying language in action by embedding psycholinguistic questions into HRI scenarios. It argues that robots can serve as controlled yet responsive partners, thereby preserving experimental robustness while allowing context to be dynamically developed. In the thesis, this framework underlies Chapter 4, where HRI is explicitly treated as a middle ground between static visual-world paradigms and fully naturalistic human–human interaction.

- Groß, A., Singh, A., Banh, N. C., Richter, B., Scharlau, I., Rohlfing, K. J., & Wrede, B. (2023). Scaffolding the human partner by contrastive guidance in an explanatory human-robot dialogue. *Frontiers in Robotics and AI*, 10.

This article demonstrates how negation and contrastive verbal guidance can scaffold the manner component of an action in an explanatory human-robot dialogue. By using instructions such as “do not push, but turn”, the study shows that contrastive guidance can operate not only at the level of locative relations but also at the functional and manner levels. In the thesis, this publication is used in Chapter 1 to motivate the functional dimension of contrast and to illustrate how contrastive contextualism extends beyond purely spatial relations.

- Singh, A., & Rohlfing, K.J. (2025). A method for studying online language-vision interplay in a human-robot interaction. *Journal of Psycholinguistic Research and Applied Psycholinguistics* (under submission).

This paper develops the methodological core for studying online language-vision processes in an interactive Human-Robot setting. It goes beyond classical visual world paradigms by replacing static displays with an unfolding task in which a robot provides explanations. The method specifies how gaze, language, and action performance can be studied in parallel, and how these streams can be aligned to capture online correspondences between linguistic guidance and visual attention. In the current work, this appears in Chapter 3 and Chapter 4, supporting the transition from treating context as static to treating it as something that is continuously developed in interaction. The proposed HRI scenario also provides a future test bed for evaluating the InterACT-R model, in particular for examining how scaffolding and fading are achieved over time.

Related peer-reviewed publications not explicitly cited in the thesis:

- Singh, A., & Rohlfing, K.J. (2022). Does contrastive attention guidance facilitate action recall? An eye-tracking study. In: Ferstl, E., Konieczny, L., von Stülpnagel, R., Beck, J., and Zacharski, L. (Eds.). *Proceedings of the Biannual Conference of the German Society for Cognitive Science (KogWis 2022)*, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany. Sep. 5th–7th 2022.
- Singh, A., & Rohlfing, K.J. (2022). Negation processing in action: How action path is interpreted in a negated context? In: Frau, F., Bischetti, L., Pompei, C., Scalingi, B., Domaneschi, F., & Bambini, V. (Eds.), *Book of Abstracts – XPRAG 2022*.

Bibliography

- Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction*, 6(1), 25–63.
- Albert, S., & de Ruiter, J. P. (2018). Repair: The interface between interaction and cognition. *Topics in Cognitive Science*, 10(2), 279–313.
- Aleven, V., & Koedinger, K. R. (2006). Design and evaluation of a tutorial interface that teaches by having students explain examples. In *Intelligent tutoring systems* (pp. 100–111). Springer.
- Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38(4), 419–439.
- Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.
- Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates.
- Anderson, J. R. (2007). How can the human mind occur in the physical universe? *Oxford University Press*.
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036–1060.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). *Cognitive tutors: Lessons learned* (Vol. 4). Taylor & Francis.
- Anderson, S. E., Chiu, E., Huette, S., & Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta Psychologica*, 137(2), 181–189.
- Au, T. K.-f., & Laframboise, D. E. (1990). Acquiring color names via linguistic contrast: The influence of contrasting terms. *Child Development*, 61(6), 1808–1823.
- Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, 120(3), 338–375.
- Bach, P., Gunter, T. C., Knoblich, G., Prinz, W., & Friederici, A. D. (2009). Understanding others' actions: The conceptual unification of functional users and objects. *Cognition*, 110(3), 349–362.
- Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, 36(2), 190.

- Bahrack, L. E., Lickliter, R., & Flom, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental Psychobiology*, 41(4), 352–363.
- Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. (2001). Intentions inferred and maintained: From acts to intentions using sensorimotor behaviors. *Intentions and Intentionality: Foundations of Social Cognition*, 37–64.
- Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(4), 723–742.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191.
- Bangerter, A., & Clark, H. H. (2004). The dynamic creation of common ground in joint reference. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 201–234). Psychology Press.
- Bar, M. (2004). Visual context facilitates object recognition. *Nature Reviews Neuroscience*, 5(8), 617–629.
- Barker, R. G. (1968). *Ecological psychology: Concepts and methods for studying the environment of human behavior*. Stanford University Press.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4), 577–660.
- Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59, 617–645.
- Barsalou, L. W., Dutriaux, L., & Scheepers, C. (2018). Moving beyond the distinction between concrete and abstract concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1752), 20170144.
- Barwise, J., & Perry, J. (1983). *Situations and attitudes*. MIT Press.
- Biederman, I. (1981). On the semantics of a glance at a scene. *Perceptual Organization*, 213, 253.
- Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing violation. *Cognitive Psychology*, 14(2), 143–177.
- Boroditsky, L. (2001). Does language shape thought?: Mandarin and english speakers' conceptions of time. *Cognitive Psychology*, 43(1), 1–22.
- Bourdieu, P. (1991). *Language and symbolic power*. Harvard University Press.
- Brennan, S. E. (2004). In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 57–78). Psychology Press.
- Bruner, J. S. (1978). The role of dialogue in language acquisition. In A. Sinclair, R. J. Jarvella, & W. J. M. Levelt (Eds.), *The child's conception of language* (pp. 241–256). Springer.

- Buschmeier, H., Buhl, H. M., Kern, F., Grimminger, A., Beierling, H., Fisher, J., Groß, A., Horwath, I., Klowait, N., Lazarov, S., Lenke, M., Lohmer, V., Rohlfing, K., Scharlau, I., Singh, A., Terfloth, L., Vollmer, A.-L., Wang, Y., Wilmes, A., & Wrede, B. (2025). Forms of understanding for XAI-explanations. *Cognitive Systems Research*, 94, 101419.
- Buxbaum, L. J. (2001). Knowledge of tool manipulation in apraxia: Functional manipulation knowledge is distinct from object structural knowledge. *Cognitive Neuropsychology*, 18(5), 489–514.
- Carlson-Radvansky, L. A. (1999). Reference frames, relations, and objects: Orientation and the structure of space. *Psychonomic Bulletin & Review*, 6(2), 286–294.
- Carruthers, P. (2006). *The architecture of the mind*. Oxford University Press.
- Chemero, A. (2009). *Radical embodied cognitive science*. MIT press.
- Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1(1), 73–105.
- Chi, M. T., Siler, S. A., & Jeong, H. (2001). Socratic tutorial dialogue. *Cognition and Instruction*, 19(3), 271–336.
- Choi, S., & Bowerman, M. (2006). Conceptual development in crosslinguistic perspective: The case of motion. *Language in Cognitive Development: The Emergence of the Conceptual Categories of Space and Time*, 151–186.
- Choi, S., McDonough, L., Bowerman, M., & Mandler, J. M. (1999). Early sensitivity to language-specific spatial categories in english and korean. *Cognitive Development*, 14(2), 241–268.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19.
- Clark, H. H. (1996). *Using language*. Cambridge University Press.
- Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association.
- Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, 3(3), 472–517.
- Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1), 62–81.
- Clarke, A. D. F., Elsner, M., & Rohde, H. (2013). Where’s wally: The influence of visual salience on referring expression generation. *Frontiers in Psychology*, 4, 329.
- Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*, 453, 494.
- Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, 17(4), 297–338.
- Cooper, R. M. (1974). Visual control of eye movements. *Perception & Psychophysics*, 16(2), 175–180.

- Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278.
- Cosper, P. (2025). Turn-taking in human communication: Origins, development, and variation. *Annual Review of Linguistics*, 11.
- Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, 13(4), 148–153.
- De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences*, 6(4), 485–507.
- De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, 14(10), 441–447.
- de Vega, M., Morera, Y., León, I., Beltrán, D., Casado, P., & Martín-Loeches, M. (2016). Sentential negation might share neurophysiological mechanisms with action inhibition. evidence from frontal theta rhythm. *Journal of Neuroscience*, 36(22), 6002–6010.
- Dennett, D. C. (1987). *The intentional stance*. MIT Press.
- Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1), 193–222.
- Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gidman, K. H., Enfield, N. J., et al. (2015). Universal principles in the repair of communication problems. *PLOS ONE*, 10(9), e0136100.
- Dourish, P. (2001). *Where the action is: The foundations of embodied interaction*. MIT Press.
- Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8(1), 19–30.
- Draschkow, D., Kallmayer, M., & Nobre, A. C. (2021). When natural behavior engages working memory. *Current Biology*, 31(4), 869–874.e5.
- Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, 92(1), 53–78.
- Duranti, A., & Goodwin, C. (1992). *Rethinking context: Language as an interactive phenomenon*. Cambridge University Press.
- Emery, N. J. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6), 581–604.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886.
- Epstein, Y., Duberstein, A., Jokinen, J., & Parsons, S. (2017). Trust and the social robot: Psychological assessment and mechanistic modeling. *IEEE Transactions on Human-Machine Systems*, 47(4), 1–10.
- Ferreira, F. (2005). Psycholinguistics, formal grammars, and cognitive science. In *The linguistic review* (pp. 365–380, Vol. 22).

- Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *Quarterly Journal of Experimental Psychology*, 61(6), 825–850.
- Fodor, J. A. (1975). *The language of thought*. Harvard University Press.
- Fodor, J. A. (1983). *The modularity of mind*. MIT press.
- Foucault, M. (1980). *Power/knowledge: Selected interviews and other writings, 1972-1977*. Pantheon.
- Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from pirahã language and cognition. *Cognition*, 108(3), 819–824.
- Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, 108(3), 316–355.
- Garfinkel, H. (1967). *Studies in ethnomethodology*. Prentice-Hall.
- Garner, W. R. (1974). *The processing of information and structure*. Erlbaum.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155–170.
- Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195–235). MIT Press.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.
- Gleitman, L. (1990). Structural sources of verb meanings. *Language Acquisition*, 1(1), 3–55.
- Gleitman, L., & Papafragou, A. (2005). Language and thought. *The Cambridge Handbook of Thinking and Reasoning*, 633–661.
- Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, 20(1), 1–19.
- Goffman, E. (1967). *Interaction ritual: Essays on face-to-face behavior*. Aldine Transaction.
- Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harvard University Press.
- Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, 69(2), 133–149.
- Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1), 3–28.
- Golinkoff, R. M., Hirsh-Pasek, K., Mervis, C. B., Frawley, W. B., & Parillo, M. (1995). Lexical principles can be extended to the acquisition of verbs. In *Beyond names*

- for things: Young children's acquisition of verbs (pp. 185–221). Lawrence Erlbaum Associates.
- Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10), 1489–1522.
- Goodwin, C. (2003). Pointing as situated practice. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 217–241). Lawrence Erlbaum Associates.
- Green, C., & Hummel, J. E. (2006). Functional constraints on the grouping of conceptual representations. *Journal of Experimental Psychology: General*, 135(2), 135–156.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (pp. 41–58, Vol. 3). Academic Press.
- Groß, A., Singh, A., Banh, N., Richter, B., Scharlau, I., Rohlfing, K., & Wrede, B. (2023). Scaffolding the human partner by contrastive guidance in an explanatory human-robot dialogue. *Frontiers in Robotics and AI*, 10.
- Gunning, D. (2017). Explainable artificial intelligence (xai). *DARPA/I2O*.
- Hafri, A., & Firestone, C. (2021). The perception of relations. *Trends in Cognitive Sciences*, 25(6), 475–492.
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335–346.
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2), 243–259.
- Henderson, J. M., & Ferreira, F. (Eds.). (2004). *The interface of language, vision, and action: Eye movements and the visual world*. Psychology Press.
- Henderson, J. M., & Hollingworth, A. (1999). Eye movements during scene viewing: An overview. *Eye Guidance in Reading and Scene Perception*, 269–293.
- Herskovits, A. (1986). *Language and spatial cognition: An interdisciplinary study of the prepositions in english*. Cambridge University Press.
- Hespos, S. J., Grossman, S. R., & Saylor, M. M. (2010). Infants' ability to parse continuous actions: Further evidence. *Neural Networks*, 23(8–9), 1026–1032.
- Hespos, S. J., Saylor, M. M., & Grossman, S. R. (2009). Infants' ability to parse continuous actions. *Developmental Psychology*, 45(2), 575–585.
- Hofstadter, D. R. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. Basic books.
- Horn, L. R. (1989). *A natural history of negation*. University of Chicago Press.
- Huetting, F., Rommers, J., & Meyer, A. S. (2011). Eye movements and spoken language comprehension: Insights from visual world studies. *Acta Psychologica*, 137(2), 151–171.
- Hunt, R. R., & Worthen, J. B. (2006). *Distinctiveness and memory*. Oxford University Press.
- Hutchins, E. (1995). *Cognition in the wild*. MIT press.

- Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), 1254–1259.
- Jackendoff, R. (1983). *Semantics and cognition*. MIT Press.
- Jung, M. F., & Hinds, P. J. (2016). Robots in the wild: A time for more robust theories. In *Human-robot interaction: Strategies and prescriptions for ai design*.
- Just, M. A., & Carpenter, P. A. (1971). The comprehension of negatives. *Journal of Verbal Learning and Verbal Behavior*, 10(3), 244–253.
- Kahn Jr, P. H., Ishiguro, H., Friedman, B., & Kanda, T. (2006). What is a human? toward psychological benchmarks in the field of human-robot interaction. *RO-MAN 2006: The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 364–371.
- Kaup, B., & Dudschig, C. (2020a). Understanding negation: Issues in the processing of negation. In V. Déprez & M. T. Espinal (Eds.), *The oxford handbook of negation*. Oxford University Press.
- Kaup, B. (1997). The processing of negatives. In *Proceedings of the 19th annual conference of the cognitive science society*. Lawrence Erlbaum Associates.
- Kaup, B., & Dudschig, C. (2020b). Understanding negation: Issues in the processing of negation. In V. Déprez & M. T. Espinal (Eds.), *The oxford handbook of negation*. Oxford University Press.
- Kaup, B., Günther, F., & Dudschig, C. (2025). The interplay of modal and amodal processes in language comprehension at the word level and beyond. In K. D. Federmeier & M. Troyer (Eds.), *Psychology of learning and motivation* (pp. 97–154, Vol. 83). Academic Press.
- Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Negation and its alternatives in a situation model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(4), 848.
- Kaup, B., Ulrich, R., Bausenhart, K. M., et al. (2024). Modal and amodal cognition: An overarching principle in various domains of psychology. *Psychological Research*, 88, 307–337.
- Kaup, B., Yaxley, R. H., Madden, C. J., Zwaan, R. A., & Lüdtke, J. (2007). The experiential simulations of negations. *Quarterly Journal of Experimental Psychology*, 60(7), 976–990.
- Kelemen, D. (1999). Function, goals and intention: Children’s teleological reasoning about objects. *Trends in Cognitive Sciences*, 3(12), 461–468.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.
- Kendrick, K. H. (2015). The selectivity of repair: An initiation-outcome bias in dyadic conversation. *Research on Language and Social Interaction*, 48(1), 78–96.
- Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18(4), 513–549.

- Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19(3), 239–264.
- Koffka, K. (1935). *Principles of gestalt psychology*. Harcourt, Brace; Company.
- Köller, W. (2016). *Formen und funktionen der negation: Untersuchungen zu den erscheinungsweisen einer sprachuniversalie*. Walter de Gruyter GmbH & Co KG.
- Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, 12(2), 72–79.
- Lajoie, S. P. (2005). Extending the scaffolding metaphor. *Instructional Science*, 33, 541–557.
- Lakusta, L., & Landau, B. (2005). Starting at the end: The importance of goals in spatial language. *Cognition*, 96(1), 1–33.
- Landau, B., & Jackendoff, R. (1993). “what” and “where” in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2), 217–238.
- Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford University Press.
- Lebiere, C., & Anderson, J. R. (1998). Cognitive arithmetic. In J. R. Anderson & C. Lebiere (Eds.), *The atomic components of thought* (pp. 297–342). Lawrence Erlbaum Associates.
- Levinson, S. C. (1983). *Pragmatics*. Cambridge university press.
- Levinson, S. C. (2016). Turn-taking in human communication—origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1), 6–14.
- Loewenstein, J., & Gentner, D. (2005). Spatial language and spatial representation: A cross-linguistic comparison. *Language*, 81(1), 128–157.
- Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 4(4), 565–572.
- Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2), 130–141.
- Lovett, M. C., Reder, L. M., & Lebiere, C. (2000). Modeling individual differences: An act-r perspective. *International Journal of Human-Computer Studies*, 53(4), 605–607.
- Luce, R. D. (1959). Individual choice behavior: A theoretical analysis.
- Lupyan, G. (2005). Language as a tool for cognition. *Behavioral and Brain Sciences*, 28(4), 503–504.
- Lupyan, G. (2012). Language is more than just a communicative system. *Cognitive Science*, 36(4), 569–574.
- Madden, C. J., & Zwaan, R. A. (2003). Grammatical aspect limits motion simulation. *Memory & Cognition*, 31(5), 663–672.
- Mandler, J. M. (1984). *Stories, scripts, and scenes: Aspects of schema theory*. Erlbaum.
- Marge, M., Espy-Wilson, C., Ward, N. G., Alwan, A., Artstein, R., Bansal, M., Blankenship, G., Chai, J., Daume III, H., Deane, P., et al. (2020). Spoken language

- interaction with robots: Research issues and recommendations. *arXiv preprint arXiv:2011.05533*.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Freeman.
- Marton, F., & Booth, S. (1997). *Learning and awareness*. Erlbaum.
- Mazza, V., Turatto, M., & Caramazza, A. (2009). Figure-ground segmentation helps handle multiple objects in visual short-term memory. *Journal of Vision*, 9(1), 1–15.
- McCarthy, J. (1993). Notes on formalizing context. *IJCAI*, 93, 555–560.
- McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, 1(1), 11–38.
- Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100(2), 254–278.
- Merleau-Ponty, M. (1962). *Phenomenology of perception* (C. Smith, Trans.). Routledge & Kegan Paul.
- Milgram, S. (1992). The individual in a social world: Essays and experiments. In *The individual in a social world*. McGraw-Hill.
- Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211–277). McGraw-Hill.
- Muir, B. M., & Moray, N. P. (1999). Trust in human-robot interaction. *Human-Computer Interaction: An Empirical Research Journal*, 4(1), 1–22.
- Murphy, G. L. (2002). *The big book of concepts*. The MIT Press.
- Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education*, 10(1), 98–129.
- Newell, A. (1980). Physical symbol systems. *Cognitive Science*, 4(2), 135–183.
- Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.
- Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 113–126.
- Nordmeyer, A. E., & Frank, M. C. (2014). The processing of negation in context. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.
- Norman, D. A. (1988). *The psychology of everyday things*. Basic Books.
- Ockham, W. o. (1974). *Summa logicae* (P. Boehner, G. Gál, & S. F. Brown, Eds.; Vol. 1). The Franciscan Institute.
- Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12), 520–527.
- Orenes, I., Beltrán, D., & Santamaría, C. (2014). It’s not that easy to say no: Processing the empty set. *Acta Psychologica*, 149, 106–110.
- Osiurak, F., Jarry, C., & Le Gall, D. (2010). Psychological bases of technical reasoning. *Journal of Experimental Psychology: General*, 139(3), 417–439.

- Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. *Cognitive science*, 34(6), 1064–1092.
- Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Naming the event: Effects of language on perception and memory for motion events. *Cognition*, 108(1), 155–184.
- Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The Journal of the Learning Sciences*, 13(3), 423–451.
- Pecher, D., & Zwaan, R. A. (Eds.). (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking*. Cambridge University Press.
- Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2), 169–190.
- Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. MIT Press.
- Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2016). Faster teaching via pomdp planning. *Cognitive Science*, 40(6), 1290–1332.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), 59.
- Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. MIT Press.
- Regier, T., & Zheng, M. (2007). Attention to endpoints: A cross-linguistic constraint on spatial meaning. *Cognitive Science*, 31(4), 705–735.
- Repp, S. (2010). Defining ‘contrast’ as an information-structural notion in grammar. *Lingua*, 120(6), 1333–1345.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.
- Reynolds, R., Zacks, J. M., & Braver, T. S. (2009). A computational model of attention for visually guided learning. *Cognitive Science*, 33(6), 1201–1228.
- Riddoch, M. J., Humphreys, G. W., Edwards, M. G., Baker, T., & Willson, K. (2003). Affordances and action: The role of action relations in object identification. *Cognitive Neuropsychology*, 20(3-6), 413–451.
- Riek, L. D. (2012). Wizard of oz studies in hri: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1), 119–136.
- Rogers, Y., Sharp, H., & Preece, J. (2011). *Interaction design: Beyond human-computer interaction*.
- Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., Esposito, E., Grimminger, A., Hammer, B., Häb-Umbach, R., Horwath, I., Hüllermeier, E., Kern, F., Kopp, S., Thommes, K., Ngonga Ngomo, A.-C., Schulte, C., Wachsmuth, H., Wagner, P., & Wrede, B. (2021). Explanation as a social

- practice: Toward a conceptual framework for the social design of ai systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13(3), 717–728.
- Rohlfing, K. J., Främling, K., Lim, B., Alpsancar, S., & Thommes, K. (Eds.). (2026). *Social explainable ai: Communications of nii shonan meetings*. Springer Singapore.
- Rohlfing, K. J., Rehm, M., & Goecke, K. U. (2003). Situatedness: The interplay between context(s) and situation. *Journal of Cognition and Culture*, 3(2), 132–156.
- Romero, O. J., Zimmerman, J., Steinfeld, A., & Tomasic, A. (2024). Synergistic integration of large language models and cognitive architectures for robust ai: An exploratory analysis. *Proceedings of the AAAI Symposium Series*, 2(1), 396–405.
- Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1(1), 75–116.
- Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3), 192–233.
- Roseberry, S., Hirsh-Pasek, K., Parish-Morris, J., & Golinkoff, R. M. (2009). Live action: Can young children learn verbs from video? *Child Development*, 80(5), 1360–1375.
- Rubin, E. (1915). *Synsoplevede figurer: Studier i psykologisk analyse* [English translation: Visually Perceived Figures (1921)]. Gyldendal.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.
- Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1), 13–24.
- Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Erlbaum.
- Schegloff, E. A. (1992). In another context. In A. Duranti & C. Goodwin (Eds.), *Rethinking context: Language as an interactive phenomenon* (pp. 191–227). Cambridge University Press.
- Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2), 361–382.
- Schwartz, D. L., & Black, J. B. (1996). Mental imagery in mental rotation and gear rotation tasks. *Journal of Experimental Psychology: General*, 125(2), 211–244.
- Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., & Sandini, G. (2012). Measuring human-robot interaction through motor resonance. *International Journal of Social Robotics*, 4(3), 223–234.
- Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.
- Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2), 70–76.

- Sera, M. D., Elieff, C., Forbes, J., Burch, M. C., Rodríguez, W., & Dubois, D. P. (2002). When language affects cognition and when it does not: An analysis of grammatical gender and classification. *Journal of Experimental Psychology: General*, 131(3), 377–397.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423.
- Singh, A., & Rohlfing, K. J. (2023). Contrastiveness in the context of action demonstration: An eye-tracking study on its effects on action perception and action recall. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45.
- Singh, A., & Rohlfing, K. J. (2024). Coupling of task and partner model: Investigating the intra-individual variability in gaze during human–robot explanatory dialogue. *Proceedings of the 26th ACM International Conference on Multimodal Interaction (ICMI '24)*.
- Singh, A., & Rohlfing, K. J. (2025a). Contrastive verbal guidance: A beneficial context for attention to events and their memory? *Cognitive Science*, 49, e70096.
- Singh, A., & Rohlfing, K. J. (2025b). A method for studying online language–vision interplay in a human–robot interaction [under submission]. *Journal of Psycholinguistic Research and Applied Psycholinguistics*.
- Singh, A., & Rohlfing, K. J. (2025c). Embedding psycholinguistics: An interactive framework for studying language in action. *Proceedings of the 6th Biannual Conference of the German Society for Cognitive Science (KogWis 2025)*.
- Singh, A., Wrede, B., Richter, B., Groß, A., & Rohlfing, K. J. (2025). Manners matter: Action history guides attention and repair choices during interaction. *IEEE International Conference on Development and Learning (ICDL)*.
- Sitnikova, T., Kuperberg, G., & Holcomb, P. J. (2003). Perception of real-world scenes: Erp evidence for the time course of semantic processing. *Neuropsychologia*, 41(9), 1122–1137.
- Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67, 101178.
- Skanze, G. (2021). Conversational ai: The science behind the alexa prize. *AI Magazine*, 42(1), 20–34.
- Slobin, D. I. (1996). *From "thought and language" to "thinking for speaking"* [In J. J. Gumperz S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70-96)]. Cambridge University Press.
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Blackwell.
- Spivey, M. (2007). *The continuity of mind*. Oxford University Press.
- Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.
- Taatgen, N. A., & Lee, F. J. (2013). Act-r as a cognitive modeling toolkit: Past, present, and future. *Topics in Cognitive Science*, 5(3), 555–563.
- Talmy, L. (1975). Semantics and syntax of motion. In J. P. Kimball (Ed.), *Syntax and semantics* (pp. 181–238, Vol. 4). Academic Press.
- Talmy, L. (1985a). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description* (pp. 57–149, Vol. 3). Cambridge University Press.
- Talmy, L. (1985b). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description* (pp. 57–149, Vol. 3). Cambridge University Press.
- Talmy, L. (2000). *Toward a cognitive semantics* (Vol. 1). MIT Press.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634.
- Tawfik, A. A., Law, V., Ge, X., Xing, W., & Kim, K. (2018). The effect of sustained vs. faded scaffolding on students’ argumentation in ill-structured problem solving. *Computers in Human Behavior*, 87, 436–449.
- Thomaz, A. L., & Breazeal, C. (2006). Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. *AAAI*, 6, 1000–1005.
- Tomasello, M. (2008). *Origins of human communication*. MIT press.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327–352.
- Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). MIT Press.
- van Ede, F., & Nobre, A. C. (2023). Turning attention inside out: How working memory serves behavior. *Annual Review of Psychology*, 74, 137–165.
- VanLehn, K. (1991). Mind bugs: The origins of procedural misconceptions. *MIT Press*.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221.
- Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press.
- Verhagen, A. (2005). *Constructions of intersubjectivity: Discourse, syntax, and cognition*. Oxford University Press.
- Viethen, J., & Dale, R. (2008). The use of spatial relations in referring expression generation. *Proceedings of the Fifth International Natural Language Generation Conference (INLG 2008)*, 59–67.
- Vigliocco, G., Vinson, D. P., Paganelli, F., & Dworzynski, K. (2005). Language and the mapping of concepts. *Cognitive Science*, 29(4), 501–533.

- Vollmer, A.-L., Mühlig, M., Steil, J. J., Pitsch, K., & Rohlfing, K. J. (2010). Robots as social actors? *Proceedings of the 19th International Symposium on Robot and Human Interactive Communication*, 688–694.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Wason, P. C. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior*, 4(1), 7–11.
- Wertheimer, M. (1923). Untersuchungen zur lehre von der gestalt. ii. *Psychologische Forschung*, 4(1), 301–350.
- Winograd, T., & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. Ablex Publishing.
- Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3), 253–265.
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89–100.
- Wu, C.-C., Wick, F. A., & Pomplun, M. (2014). Interactions between conceptual knowledge and visual attention. *Visual Cognition*, 22(6), 811–841.
- Zacks, J. M., & Tversky, B. (2001). Perception and memory of events. *Journal of Experimental Psychology: General*, 130(1), 29–47.
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2), 162–185.

Cognitive Science 49 (2025) e70096

© 2025 The Author(s). *Cognitive Science* published by Wiley Periodicals LLC on behalf of Cognitive Science Society (CSS).

ISSN: 1551-6709 online

DOI: 10.1111/cogs.70096

Contrastive Verbal Guidance: A Beneficial Context for Attention To Events and Their Memory?

Amit Singh,  Katharina J. Rohlfing 

Psycholinguistics, Paderborn University

Received 27 August 2024; received in revised form 29 June 2025; accepted 9 July 2025

Abstract

Research suggests that presenting an action via multimodal stimulation (verbal and visual) enhances its perception. To highlight this, in most studies, assertive instructions are generally presented before the occurrence of the visual subevent(s). However, verbal instructions need not always be assertive; they can also include negation to contrast the present event with a prior one, thereby facilitating processing—a phenomenon known as contextual facilitation. In our study, we investigated whether using negation to guide an action sequence facilitates action perception, particularly when two consecutive subactions contrast with each other. Stimuli from previous studies on action demonstration were used to create (non)contrastive actions, that is, a ball following noncontrastive and identical (Over–Over or Under–Under) versus contrastive and opposite paths (Over–Under or Under–Over) before terminating at a goal location. In Experiment 1, either an assertive or a negative instruction was provided as verbal guidance before onset of each path. Analyzing data from 35 participants, we found that, whereas assertive instructions facilitate overall action recall, negating the later path for contrastive actions is equally facilitative. Given that action goal is the most salient aspect in event memory due to goal-path bias in attention, a second experiment was conducted to test the effect of multimodal synchrony on goal attention and action memory. Experiment 2 revealed that when instructions overlap with actions, they become more tailored—assertive instructions effectively guide noncontrastive actions, while assertive–negative instruction particularly guides contrastive actions. Both studies suggest that increased attention to the goal leads to coarser perception of midevents, with action-instruction synchrony modulating goal bias in real-time event apprehension to serve distinct purposes for action conceptualization. Whereas presenting instructions before subactions attenuates goal attention,

Correspondence should be sent to Amit Singh, Psycholinguistics, Faculty of Arts and Humanities, Paderborn University, Zukunftsmeile 2, 33102 Paderborn, Germany. E-mail: amit.singh@uni-paderborn.de

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

overlapping instructions increase goal attention and reveal the selective roles of assertive and negative instructions in guiding contrastive and noncontrastive actions.

Keywords: Event cognition; Contrastive guidance; Attention; Negation; Eye tracking

1. Introduction

Negation is ubiquitous in all the languages which enables us to talk about nonexistence, express rejection, denial, and impose prohibition (Choi, 1988; Horn, 2001). For example, an utterance “It will not rain today” let us express our belief that a particular event (namely: rain) will not occur. Given sufficient context, a listener might also be convinced to make an inference that it had been raining for the past few days but not today. In this way, negation allows speakers to convey something which counters listener’s expectation (Choi, 1988; Drozd, 1995; Dale & Duran, 2011; Glenberg, Robertson, Jansen, & Johnson-Glenberg, 1999; Lüdtke & Kaup, 2006; Orenes, Beltrán, & Santamaría, 2014; Stern, 1931; Wason, 1965). Thus, the use of negation seems to be highly context-dependent (Albu, Tsaregorodtseva, & Kaup, 2021; Kaup & Dudschig, 2020), where one aims to contrast the present with the past event(s) (Glenberg et al., 1999; Horn, 2001). Although negation is among the most frequently used constructs in a language, it is often associated with difficulty in processing, false memory, and slower retrieval of concepts within its scope compared to affirmative counterparts (Kaup, Lüdtke, & Zwaan, 2006; Kaup, Zwaan, & Lüdtke, 2007; Mayo, Schul, & Rosenthal, 2014). In context of action performance, negation has been shown to cause inhibition at cognitive and neural level (Beltrán, Liu, & de Vega, 2021; de Vega et al., 2016). While studies have reported processing difficulty for sentential negations (Carpenter & Just, 1975; Clark & Chase, 1972), in the event cognition literature, negation remains an underexplored domain (Dudschig, Kaup, Liu, & Schwab, 2021), yet it could reveal promising consequences for action processing (e.g., Groß et al., 2023). In event processing research, it is known that the previous subevent acts as a context for the upcoming subevent, and in its entirety, event memory is more robust when the subevents are connected at some spatio-temporal dimension (Zacks et al., 2001; Zacks & Swallow, 2007; Zacks, Swallow, Vettel, & McAvoy, 2006). The limited capacity of our attentional system highlights such importance of the coherence, that is, improved learning when modalities are unified across multiple levels—at verbal and visual (Bahrick, Lickliter, & Flom, 2004; Brand & Tapscott, 2007). It has also been found that participants often use negation to contrast and refer back to some previous subevent(s), such that something occurred previously but not this time (e.g., Tian, Ferguson, & Breheny, 2016). For such events, a verbal guidance which refers back to the previous subevent, while contrasting it with the present subevent, is more likely to facilitate a deeper connection at multimodal level. Linguistic negation is one such means which derives the meaning of an entity under its scope by referring to some alternate state of affair (Kaup et al., 2007; Talmy, 1975, 1985), thereby creating a rich pragmatic effect (Tian & Breheny, 2016). This unique feature highlights negation as a context-generating constituent that bridges the present and the previous subevents (Zacks et al., 2006). For example, in a recent study, participants were shown a negative sentence either with or without a context contrasting an expected situation—for

instance, *contrary to expectation*, *John did not iron his shirt* (Albu et al., 2021). In a probe recognition task, it was found that participants responded faster to a sentence when the expectation was contrasted with actual state of affair. Here, the later subevent “did not iron” might have been interpreted in light of the prior subevent “did iron,” which acted as a context, and thus eased its processing. It has also been shown that participants are more likely to negate a stimuli depicting an absence of a feature when surrounding stimuli already possess that particular feature (Nordmeyer & Frank, 2023; Wason, 1965). This suggests a situated use of negation, where an event in hand can be defined by the absence of a prior event or one occurring in close vicinity, especially when they are temporally or spatially bound. Based on these findings, it has been suggested that context can mitigate processing cost associated with negations (Lüdtke & Kaup, 2006). Nevertheless, the effect of contextual facilitation on negation has been studied mostly for linguistic tasks and has rarely been examined in the context of actions, which inherently incorporate temporality, that is, events that unfold continuously over time (Huff, Papenmeier, & Zacks, 2012; Zacks et al., 2001). The majority of studies in this area focus on how access of a concept is eased, when the surrounding context is static, or provided before a sentence. When both sentence processing and action performance are brought together, a close interplay between linguistic processes and event conceptualization is observed (Altmann & Kamide, 2004). Crucial to this interplay, one perspective argues that language influences even the very early event perception, guiding attention through its innate semantic structures, which becomes a part of procedural knowledge through language use (Levelt, 1993). Conversely, another view suggests that early attentional processes are guided by domain-general mechanisms, wherein event conceptualization at the perceptual stage remains unaffected by the syntacto-semantic structure of language and is only influenced at later stages (Griffin & Bock, 2000). Despite this temporal distinction, both perspectives agree that language directs attention to event structures acting as a spotlight (Wolff & Holmes, 2011). To capture these concurrent processes, online methods like eye tracking have been employed. For instance, studies particularly in motion event show that eye movements are consistently drawn more toward the event goal (path) than other event components, such as manner, regardless of whether linguistic or nonlinguistic processes are involved (Papafragou, 2010; Zacks & Swallow, 2007). From a developmental perspective, several lines of evidence indicate that infants can register predictability structure within human activity in relation to actors’ goals, as opposed to just specific surface motion patterns actors adopt to pursue those goals (Kosie & Baldwin, 2021; Olofson & Baldwin, 2011; Woodward, 1998). They can also organize their processing of goal-directed activity around that predictability structure by, for example, directing gaze to an actor’s predicted goal object in advance to the actor’s actual contact with that object (Ambrosini et al., 2013; Kanakogi & Itakura, 2011; Monroy, Gerson, & Hunnius, 2017). A common explanation for such behavior is that observing an action inherently engages the observer’s own motor system, allowing them to simulate the action and anticipate the agent’s goal (Gredebäck & Falck-Ytter, 2015; Rizzolatti & Craighero, 2004). From a social interactional perspective, prevailing debates on how actions are learned have focused on whether, and to what extent, sensorimotor processes involved in goal anticipation are influenced by top-down linguistic and semantic structures (Griffin & Bock, 2000; Papafragou, Hulbert, & Trueswell, 2008; Wolff & Holmes, 2011). This issue was approached

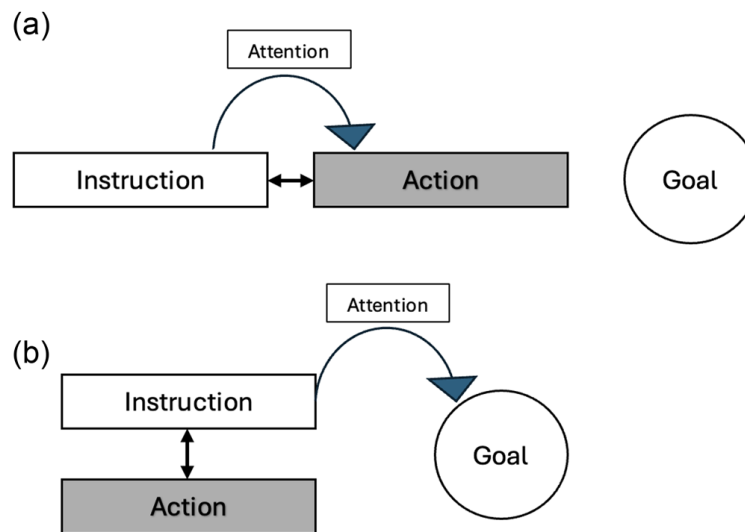


Fig. 1. Two forms of multimodal synchrony: (a) sequential and (b) simultaneous. In the sequential presentation (Experiment 1), attention is directed toward the immediate action (path) as the instruction precedes and introduces the action. In the simultaneous presentation (Experiment 2), overlapping instruction and action are expected to shift attention directly to the action goal due to ever-present goal bias in event perception.

by investigating processes related to language-mediated sensorimotor behavior, specifically testing whether language can operate on sensorimotor processes to influence event and action apprehension (Holmes & Wolff, 2010; Wolff & Holmes, 2011). The mechanisms through which language (verbal behavior) may meddle with actions can also be sensitive to how the modalities are aligned (Brand & Tapscott, 2007; Bahrack et al., 2004). In situations where language is an ostensive cue and acts as a spotlight, directing attention to the immediate subaction may mitigate the goal bias arising from inherent sensorimotor processes. A sensitivity to ostensive cues alerts learners that there is an opportunity to learn, and accordingly, important aspects are communicated (Csibra & Gergely, 2009). This is a well-documented phenomenon; for example, Nomikou and Rohlfing (2011) found that mothers used verbal instructions prior to a particular subaction to introduce and highlight it, thereby making those subactions salient in relation to the action goal (see also, Sciutti, Lohan, Gredebäck, Koch, & Rohlfing, 2016). In an alternate view, early learning has also been shown to be shaped by co-occurrence of visual and verbal information, such as, through modality overlapping (Bahrack & Lickliter, 2000, 2002). When modalities overlap, language creates sensorimotor redundancy that facilitates the detection of amodal properties of the stimulus (Bahrack et al., 2004). The two ways of multimodal correspondence between language and action may allude to different forms of instruction-action synchrony. Whereas presenting instructions prior to a subaction may subserve the learning of fine-grained aspects of an action by spotlighting elements that might otherwise go unnoticed, overlapping modalities may instead lay the foundation for action structuring through goal-directed component. Hence, considering the possible ways in which language may structure action perception, we conducted two eye tracking experiments to test two forms of instruction-action synchrony: presenting instructions either before or concurrently with actions (Fig. 1). The primary objective of this study was to ground negation

in an action performance task to determine whether negation, as a form of contrast, provides plausible context for action understanding. Subsequently, we aim to examine whether, and to what extent, action contexts are sensitive to synchrony of the modalities involved, in light of the ever-present goal bias in event memory. These findings can have implications for effective verbal guidance in many instructional settings, such as learning or working together.

1.1. Grounding negation in action performance

Although previous studies have provided limited focus on the interplay between negation and action, they are nonetheless informative about the underlying cognitive mechanisms that our investigation can capitalize upon. Recent studies have found that negation leads to action inhibition (Beltrán et al., 2021; de Vega et al., 2016), for which the effect could go back to alternate state of affair being activated before the actual state of affair (Kaup et al., 2006; Kaup et al., 2007). This proposition is made in early studies showing that a two-step mental simulation takes place as someone encounters a negated sentence. For example, in a study by Kaup et al. (2007), participants were shown a sentence, such as “There was no eagle in the sky,” and thereafter, a picture matching a negated state of affair (an eagle with stretched wings) or an actual state of affair (an eagle with folded wings). It was found that after a short Stimulus Onset Asynchrony (SOA) of 500 ms, participants responded faster to a picture matching the negated state of affair than to a picture matching an alternate state of affair. Crucially, this pattern was reversed when a sufficient time for response was provided, which presumably allowed participants to simulate the situation after reading the sentence. The results suggested that participants relied predominantly on mental simulation while processing the negative sentences. Based on findings like this, the conceptual access under negation has been shown to be governed by a two-step cognitive process, where at the first step, the activation is guided by the propositional structure, and at the later, by the situational model (Kaup et al., 2006; Kaup et al., 2007). This simulation has been shown to be constitutive of the domain general cognitive processes (Mayo et al., 2014), which directly connects to actions and embodiment accounts (Barsalou, 1999; Glenberg & Gallese, 2012). For the sake of completeness, it should be noted that other studies, however, proposed rather a single step account (e.g., Lüdtke & Kaup, 2006; Tian et al., 2016). It is suggested that negation also relates to sensorimotor system, which involves general cognitive mechanism, such as response inhibition (de Vega et al., 2016). More specifically, in an electroencephalography (EEG) study by de Vega et al. (2016), participants were asked to read an assertive or a negative sentence while doing a NoGo task; it was found that theta oscillation—a neural marker of response inhibition—was smaller in the negative than the assertive condition. Notably, the studies in favor of two-step cognitive process usually focus on picture recognition, inhibitory control, or sentential negation, where the accessibility of a critical concept is measured through recording of the reaction time and accuracy. Accordingly, a shorter latency then corresponds to easier access to the concept under consideration, and the effect of context, especially in an action performance task, is yet to be explored. If negation processing links to a domain-general processing mechanism, which relies on simulation and the contextual factors, then measuring these effects directly

for action performance, such as in an action recall task, provides a promising area of investigation.

1.2. Action and attention to the goal

We argue that negation, although mostly studied as a linguistic means, also connects deeply to actions (Bartoli et al., 2013; Tomasino, Weiss, & Fink, 2010; Tettamanti et al., 2008). For example, negation is found to be crucial in action learning during developmental stage (Bloom, 1993; Horn, 2001). It is a fundamental feature to elicit response inhibition verbally in children, for example, saying “No!” for an action not only inhibits the current response, but also acts as a scaffolding means, where correct response thereafter can implicitly or explicitly be conveyed (Bloom, 1970; Drozd, 1995). Diverting learners’ attention from a specific entity using negations and guiding it to the relevant aspects of the task is directly bound to the task goal. This method is frequently used by the caregivers toward a developing child in an interaction, be it for language acquisition (Bloom, 1970; Choi, 1988; Drozd, 1995; Hummer, Wimmer, & Antes, 1993; Tagliani, Vender, & Melloni, 2022) or action learning (Bloom, 1993; Woodward & Sommerville, 2000).

But what makes actions a special case for negations? It is now well known that human behaviors, including actions, are goal-directed (Zacks et al., 2001). This goal is maintained at a higher level of availability and is foregrounded from other constituents in an action (Ikegami, 1979; Talmy, 1985; Zacks et al., 2001). For example, infants have been shown to structure their actions in relation to goals (Kosie & Baldwin, 2021), and studies with adults report a better memory for goals than for other components—such as source—in a motion event (Ikegami, 1979; Papafragou, 2010; Talmy, 1975). In one study, participants were shown an animate object moving from a source to a goal location, after which a change in goal or source was probed. It was found that participants were more likely to report a change in the goal than the source location, suggesting that source is not encoded to the same extent as goal in a motion event (Papafragou, 2010). Additionally, in a motion event, there are two fundamental components identified: path and manner (Slobin, 1996; Talmy, 1975). In this context, the path characterizes “Where” component in a motion, while the manner specifies “How” the action should be performed. The path has also been associated with goal of a motion event, toward which the path terminates at the end (Talmy, 1985, 2000). Implementing this idea within an interactive framework, a related study using human–robot dialogue found that when either path or manner information was missing, path information was sought more quickly than manner information to complete the task at hand, suggesting that bias effect can be also elicited by the path component in actions (Singh & Rohlfing, 2024). The prioritization of goal in event cognition, while utmost beneficial for task in hand, may potentially undermine the attention on fine-grained intermediary subevent(s) within a continuous action. For instance, a multitude of studies (also in infancy research) have consistently demonstrated a coarse perception of manner than path (Bunger, Skordos, Trueswell, & Papafragou, 2021; Hespos, Grossman, and Saylor, 2010; Papafragou et al., 2008). And the manner in which an action is performed is not captured to the same level as the path or trajectory (Hespos, Saylor, and Grossman, 2009; Skordos et al., 2020). Thus, to optimally learn from action

demonstrations, it is crucial that the path and manner receive a comparable degree of attention. In such a situation, highlighting the manner in an action, and simultaneously reducing the attention from goal using a verbal means (or perhaps negation), might be beneficial for a better action conceptualization. In this respect, from a very recent study on attention, we can draw potential for negation in facilitating a more even distribution of weights across stimuli (Banh, Tünnermann, Rohlfing, & Scharlau, 2024). However, in this study, the context of negation was not systematically varied and was applied solely to static stimuli within the spatial domain, without visual contrasts. Unlike static events, a continuous unfolding action that elicits a temporally bound contrast between two subevents may offer a more suitable context for examining the effects of negation. Whereas negation has the potential to meddle with sensorimotor processes through its unique property of inhibitory processes (de Vega et al., 2016; Kaup & Dudschig, 2020), language itself has a broader propensity to affect early sensorimotor processes by spotlighting specific aspects—those that might otherwise go unattended due to a developmentally learned routines of goal prominence (Sciutti et al., 2016; Wolff & Holmes, 2011). Most of this learning takes place during early childhood through multimodal synchrony (Bahrick & Lickliter, 2000; Bahrick et al., 2004). Essentially, the tight temporal co-occurrence of verbal and action routines plays a key role in decoding the “semantics” of the action, guiding it through goal interpretation (Bahrick et al., 2004). Verbal instructions that are not strictly overlapping with actions but are instead introduced beforehand may meddle with this learned behavior, potentially reorienting or sustain attention to elements that would typically go unnoticed in continuous actions. Hence, synchronizing instructions in such a way that language acts as a spotlight—not for the most salient aspect, but on coarser, intermediate aspects of the action—may hold potential for inducing subtle but meaningful effects in instructional settings (Fig. 1).

2. Current study

The present study primarily aimed to assess the effect of context induced by negation on the perception of contrastive and noncontrastive actions. For this, we used simple action models in which a contrast is achieved on two action paths following each other. The participants had to recall these constellations. To elicit a contrast, the task would need to satisfy following requirements: (A) The contrast must occur at least at one of the components of action, namely, the path or manner, and (B) the present action component should be opposite to the prior action component on some spatio-temporal dimension. Given these requirements, the stimuli were adopted from a previous study on action learning in infants (Hespos et al., 2009; Hespos et al., 2010), where the path direction was manipulated. We selected these stimuli because they consisted of a spatial framework at the most salient dimension of path manipulation in vertical direction (e.g., Over and Under), which is salient than horizontal (Front–Back or Left–Right) due to ever present effect of gravity (Clark, Carpenter, & Just, 1973). In addition, we combined verbal guidance in assertive and negative forms to guide these paths. Overall, we expected to observe an inverse relationship between attention to the goal and action recall since this aligned with our prediction that there exists a general trade-off between attention

to the goal and memory of the mid-event constituents (e.g., manner) in the action demonstration (see Section 1.2). Our primary hypothesis relates to testing an action context that permits the use of negation. A plausible context is defined when the negation denies the expected subevent or indicates the absence of a previous subevent, such that something previously happened but not this time. We reasoned that such contexts would facilitate a rich connection between the prior and the current subevents at two levels (verbal and visual). For example, in a contrastive action, when the later path is opposite of the previous path (Over–Under), using negative guidance for the later path (e.g., NOT Over) not only de-biases from the expectation but also refers to the previous path (Over). This “back referencing” allows a connection at the verbal level as compared to the simple assertive guidance (Now Over–Now Under), which only facilitates connection at the visual level and misses the opportunity to utilize the previous context. Thus, we predicted a better memory and attenuation of attention from the goal for the contrastive action in the presence of an assertive–negative instruction. We also predicted an overall higher recall for an action when it is guided through assertions as compared to when no guidance is provided (a baseline condition), because a multimodal presentation of an action has been shown to facilitate its perception (Bahrnick & Lickliter, 2000, 2002; Brand & Tapscott, 2007). This is especially the case when an action comprises of repetitive paths or subactions without contrasting each other (noncontrastive). Hence, for noncontrastive actions, we expected higher recall and goal attenuation under assertive guidance, as negation would be implausible in the absence of visual contrast—making simple assertive guidance the most facilitative. Accordingly, we also predicted lower action recall and higher attention to the goal when a negation is used to guide a noncontrastive action, for example, a guidance like [Now Over–NOT Under] used for two upward motions. Overall, we tested all possible combinations of assertive and negative guidances on contrastive and noncontrastive actions, each comprising two paths following each other.

Previous studies have also emphasized the sensitivity of verbal–action temporal synchrony on learning and attention (Bahrnick & Lickliter, 2000, 2002; Bahrnick et al., 2004; Brand & Tapscott, 2007). Building on this, we tested our hypothesis across two experiments: instructions presented before the action (Experiment 1) and instructions presented simultaneously with the action (Experiment 2). The two forms of temporal synchrony were found to influence attention to the modalities in distinct ways. A sequential verbal–visual presentation has been frequently employed when the goal is to introduce and spotlight a particular subaction, as observed in infant–mother interactions (Nomikou & Rohlfing, 2011). Furthermore, it is suggested that unimodal stimulation promotes prolonged and sustained attention to the modality in question, making its properties more salient, as only one of the two modalities is present at any given time (Bahrnick et al., 2004). Thus, for sequential presentation, we predicted that when instructions precede actions, sustained attention on the immediate event would delay the attention on the goal (as in Fig. 1a). In contrast, when instruction and action overlap (Fig. 1b), the modalities are perceived as unitary due to a well-established phenomenon of intersensory redundancy for which our cognition is developmentally shaped (Bahrnick & Lickliter, 2000, 2002; Bahrnick et al., 2004). In such cases, the default behavior of our perceptual system is to prioritize the action goal—unless, ostensive attention is directed toward the upcoming action, as in sequential presentation (Fig. 1a). Therefore, in this condition, we predicted that the

attention would gravitate directly toward the action goal, which serves as the default locus of attention (Lakusta & Landau, 2012; Zacks et al., 2001).

3. Experiment 1: Instruction before the action

3.1. Method

3.1.1. Participants

The sample size for the study was determined by the previous eye tracking studies on the event understanding (Bunger, Skordos, Trueswell, & Papafragou, 2016; Papafragou et al., 2008). Additionally, a post-hoc power analysis was conducted on pilot study data.¹ Our goal was to achieve 0.95 power (95%) to detect a medium effect size of 0.25, with a standard alpha error probability of 0.05, which suggested a sample size of 30. Thirty-five university students (mean age = 23.90, $SD = 2.97$) who were native or fluent German speakers took part in the experiment. An informed consent was obtained prior to the data collection in accordance to the university ethical board. All the participants were recruited through classroom advertisement and were compensated either a course credit or remuneration for their participation. Data from three participants ($N = 3$) were excluded from the analysis, due to the track-loss or failing to properly follow the experiment instruction. Thus, the final sample size for this experiment was 32.

3.1.2. Stimuli

The study replicated four videos from Hespos et al. (2009); Hespos et al. (2010), depicting a ball moved from a source toward the goal. During this entire video, the ball showed two prominent changes in the path direction at the source and at the goal locations. These changes were characterized by an upward or a downward motion. A video in which the direction of the later path (at goal) contrasted the previous path (at source), we treated it as a contrastive action, for example, an upward path followed by a downward path or vice versa (Fig. 2b,d). Conversely, a video in which the direction of the later path (at goal) was a repetition of the previous path (at source), we treated it as noncontrastive action, for example, two upward or two downward paths in a sequence (Fig. 2a,c). Two types of contrastive and noncontrastive actions were created to represent the two levels of complexity. This was done to preserve some variability in the experimental items. Crucially, across these items, the nature of the action—contrastive versus noncontrastive—remained consistent, specifically involving changes in the path direction at the source and goal locations. A verbal instruction was used to guide the directions. These instructions were recorded by a female speaker in an intonation which is used while demonstrating an action. The guidance was provided in an assertive or a negative form, for instance, a guidance for the upward path direction in assertive condition was “Nach Oben!” [“Toward Over!” in English] or in negative condition “Nicht Unten!” [“Not Under!” in English]. Similarly, a downward path direction in assertive condition was “Nach Unten!” [“Toward Under!” in English] or in negative “Nicht Oben!” [“Not Over!” in English]. Crucially, the “Nach” and the “Nicht” correspond to a path preposition and negation particle,

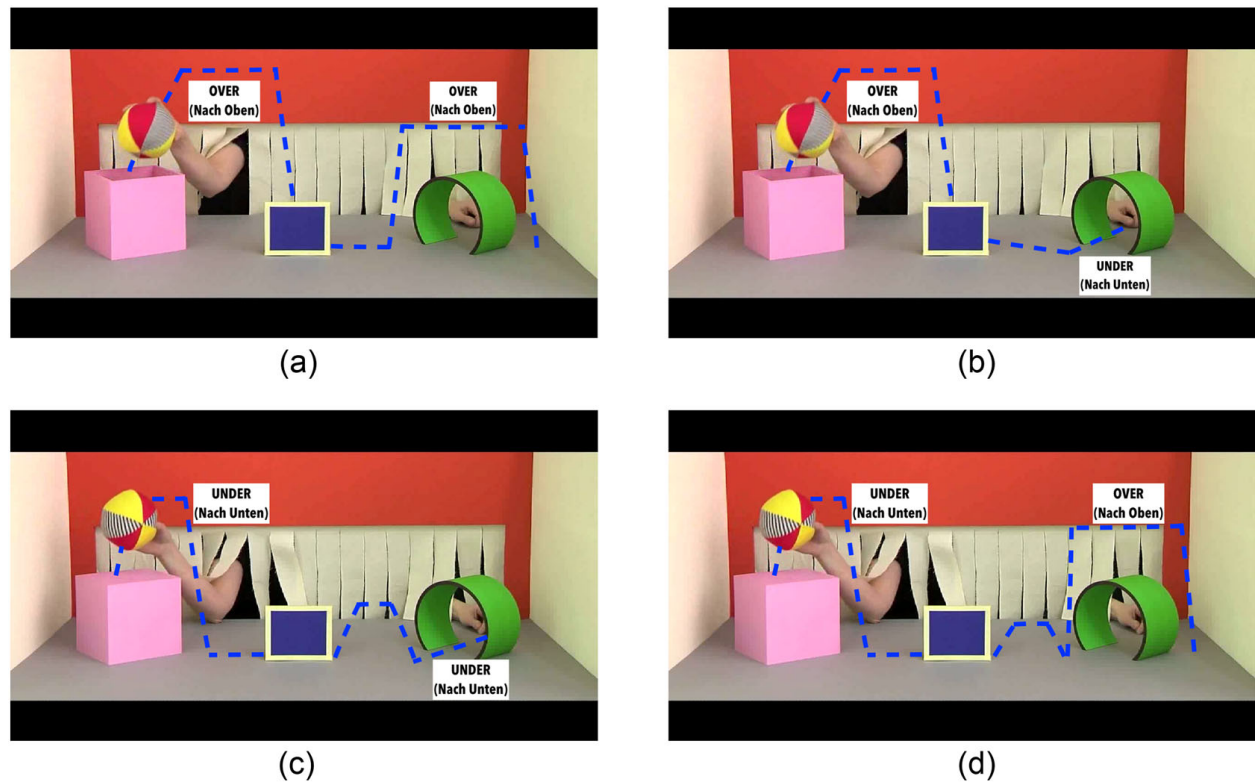


Fig. 2. Stimuli depicting the action conditions. Blue line shows the path of the ball with instruction in assertive condition.

respectively. In a motion event, the path preposition like *toward* (*Nach*) refers to an execution of an action in the direction of a reference object (Talmy, 2000). Hence, we used this preposition to highlight the direction of the path in assertive condition and to make the conditions comparable in number of words. As a result, the instruction conditions for two consecutive segments in an action were: assertive–assertive, negative–negative, assertive–negative, and negative–assertive. A video without a verbal instruction was used as the baseline. A total of 20 trials were created by combining four videos and five instruction conditions including the baseline where the instruction was absent.

3.1.3. Procedure

The experiment was designed on the 1920 X 1080 pixels monitor equipped with Tobii-Pro 120 Hz eye tracker. The size of the video was 1080 x 720 pixels. Participants were seated on a swivel chair approximately 60 cm from the screen. The objects used in the video were placed behind the participants on a small stage for the recall of the actions. A video camera was installed behind the stage to record the action performance. The participants were introduced to the experiment and the setup. They were told that they will watch 20 videos one by one in a sequence. In these videos, a ball will be moved by a person on a similar stage which is placed behind. Each video will have a verbal instruction. The task is to carefully watch the video, remember the actions shown, and then perform it on the stage. The experiment started

with a fixation followed by the first video then a probe text at the end. The probe asked the participants to turn back and perform the action shown in the video. Participants then turned back while seated, performed the action on the stage, and then returned back to the initial position to proceed with the next trial. During this entire trial, participants did not require to stand up from the chair. The videos were randomized and self-paced where participants were given unlimited time to perform the task. The next trial was initiated by pressing the space bar on the keyboard. The position of chair was marked on the ground to maintain a fixed distance between the eye tracker and the participant. Prior to the main experiment, participants were given 10 practice trials to familiarize themselves with the experimental setup. In the practice session, the objects and its path direction were different from the main experiment to avoid any carry-over effect. After the practice session, the main experiment began. The entire experiment lasted for approximately 30 min in which the eye gaze and video data were recorded.

3.1.4. Analysis

For the recall coding, the videos of action demonstration and performance were exported from the eye tracker and video recorder, respectively. For the fixation metrics, the data was extracted from the Tobii-Pro Lab software after preprocessing. All the analyses were done in R version 4.4.0 (R Core Team, 2022) and R studio v12.1.402 (Posit team, 2024).

3.1.4.1. Preprocessing of eye movement data: A polygon Area of Interest (AOI) was created for the goal object (here: Green Ring). This was based on the previous studies suggesting that the goal receives maximum attention than any other entity within an action (Bunger, Trueswell, & Papafragou, 2012; Bunger et al., 2016, 2021; Papafragou et al., 2008). To avoid any coincidence of moving ball with the goal, the goal-AOI was only active from the beginning of the trial until the ball reached the goal. This was consistent with the previous method used in event cognition to investigate the attention to the goal (Papafragou et al., 2008). Fixations were classified using Tobii-Pro-IV inbuilt filter. For a time-dependent analysis, each trial was divided into 200 ms time bins following the previous research showing that saccades approximately take 200 ms to launch (Saslow, 1967). To determine the fixation proportion in each time bin, the number of fixations falling in an AOI was divided by the total number of fixations in that bin. Participants with more than 25% track loss across all trials and trials with more than 50% track loss were excluded from the analysis.

3.1.4.2. Preprocessing of recall videos: To determine the proportion of correct recall for each participant, we compared the video presented on the eye tracker with the video recorded by the camera. Each video was segmented into subactions based on the changes in the path trajectory. These subactions were identified by counting the inflection points in the path direction, that is, the number of times the ball changes its direction in an entire video. This segmentation resulted in a total of six subactions for actions (a) and (b) (Fig. 2a,b), and eight subactions for actions (c) and (d) (Fig. 2c,d). A correct recall of the subaction was identified when the participants correctly followed the specified path direction: over or under. Conversely, if a participant performed a substitutive subaction, an entirely different subaction,

or omitted the subaction, it was classified as incorrect. Each subaction was scored 1 for correct recall and 0 for incorrect recall. Based on the transcription, we calculated the types of errors and its percentage distribution and found that omission errors occurred most frequently (51.82%), followed by substitution errors (24.82%), and other errors (23.36%). Overall, the coding of participants' subactions yielded 640 data points, among which 137 were classified as errors, resulting in an overall accuracy of 78.59%.² The total number of correct subactions were added for each video. The proportion of correct recall was calculated by normalizing the correct subactions by total subactions. For the analysis, a logistic regression model was fitted using a generalized linear mixed effects model (glmer) in R (Barr, 2008). To ensure the reliability of the coding, two student assistants coded the data for which the inter-rater reliability test was conducted (Kendall's $W = 0.78$).

4. Results

4.1. Eye movements

We analyzed the time-dependent changes to the fixation on the goal object to investigate the attention to the goal. The raw data followed a nonlinear curve, hence we used the Generalized Additive Model (GAM) to model the fixations. GAM has been shown to capture the nonlinearity by penalizing the overfitting and has previously been used to model data in visual world paradigms (Porretta, Kyröläinen, Van Rij, & Järviö, 2018; Porretta, Tucker, & Järviö, 2016; Wieling, 2018). To evaluate the assumption of nonlinearity, we initially fitted the data using both linear and nonlinear models by incorporating a smoothed (nonlinear) term. An Akaike Information Criterion (AIC) test revealed that the nonlinear model had a lower AIC score compared to the linear model (AIC difference = -8962.10), supporting the justification for nonlinearity. For our analysis, we fitted a separate GAM model for each action type. The fixations were transformed into empirical-logit to treat it as a continuous response variable (Barr, 2008). The model included parametric terms for instructions, which were treatment contrast-coded with the no-instruction condition as the baseline, along with its factor-smooth interaction with the time bin. A log-likelihood model comparison was done between a reduced and a full model to estimate the best fitting model following Barr, Levy, Scheepers, and Tily (2013). The Chi-Square test on the Maximum Likelihood (ML) scores indicated that model with smooth term was significantly better than the null model for both contrastive ($\chi^2(8.00) = 251.31, p < .001$) and noncontrastive actions ($\chi^2(8.00) = 579.66, p < .001$). The model summary is provided in Table 1, and the estimated effects at the probability scale are visualized in Fig. 3.

(A) Contrastive action: Compared to the baseline (no instruction), the parametric term (time-independent) revealed a significant main effect of instruction. The estimates indicated that fixation on the goal significantly decreased from the baseline as soon as instructions were provided across all instruction conditions: negative–negative ($\beta = -.06, p < .001$), assertive–negative ($\beta = -.05, p < .001$), negative–assertive ($\beta = -.04, p < .001$), and assertive–assertive ($\beta = -.02, p < .001$) instructions. Crucially, the time-dependent smooth terms revealed

Table 1

Experiment 1: Estimates of looks toward the goal for contrastive and noncontrastive actions

Contrastive action			
Parametric coefficients	Estimate	SE	<i>p</i>
Intercept	−3.71	.003	< .001
Assertive–Assertive	−.02	.004	< .001
Assertive–Negative	−.05	.004	< .001
Negative–Assertive	−.04	.004	< .001
Negative–Negative	−.06	.004	< .001
Smooth terms	edf	<i>F</i>	<i>p</i>
s(bin):No-instruction	8.94	358.93	< .001
s(bin):Assertive–Assertive	8.88	221.33	< .001
s(bin):Assertive–Negative	8.72	226.08	< .001
s(bin):Negative–Assertive	8.84	204.24	< .001
s(bin):Negative–Negative	8.91	215.81	< .001
Noncontrastive action			
Parametric coefficients	Estimate	SE	<i>p</i>
Intercept	−3.79	.002	< .001
Assertive–Assertive	.003	.004	.42
Assertive–Negative	.001	.004	.89
Negative–Assertive	.012	.004	.002
Negative–Negative	.02	.004	< .001
Smooth terms	edf	<i>F</i>	<i>p</i>
s(bin):No-instruction	8.95	146.01	< .001
s(bin):Assertive–Assertive	8.93	123.92	< .001
s(bin):Assertive–Negative	8.96	191.34	< .001
s(bin):Negative–Assertive	8.95	181.13	< .001
s(bin):Negative–Negative	8.94	174.47	< .001

Abbreviation: Intercept, no-instruction condition.

distinct temporal trends across all the instruction conditions. The assertive–negative condition exhibited the lowest effective degrees of freedom ($\text{edf} = 8.72$, $p < .001$), indicating a smoother trajectory over time compared to the other conditions (Fig. 3, top). To examine this time-dependent reduction in goal-looks, we conducted GAM difference smooth comparisons for all the instruction conditions against the baseline (no instruction). As revealed in the visualization, fixation on the goal in the assertive–negative condition significantly decreased in the later time period relative to the baseline.³

(B) Noncontrastive action: For the parametric term (time-independent), a significant difference in goal-looks from the baseline (no-instruction) was observed when a negative instruction was provided first, with the higher fixation from the baseline for the negative–negative instruction ($\beta = .02$, $p < .001$) and then negative–assertive ($\beta = .012$, $p = .002$). Regarding the time-dependent changes, all the instruction conditions were significantly different from

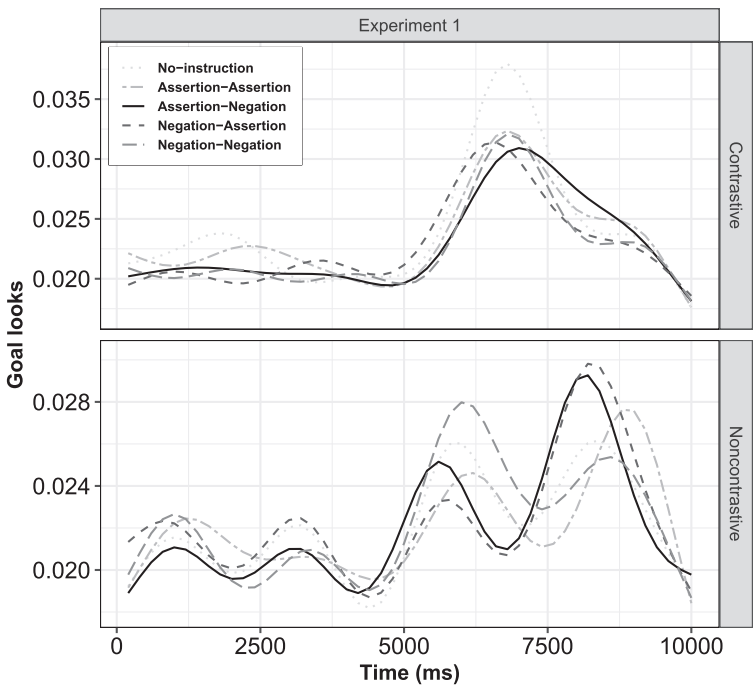


Fig. 3. Experiment 1: Estimated goal looks in contrastive (top plot) and noncontrastive action (bottom plot) conditions.

reference, with the lowest edf score for the curve of assertive–assertive condition ($\text{edf} = 8.93$, $p < .001$) (Fig. 3, bottom). To test for the time-related changes to the goal for each instruction condition from the baseline, we conducted a difference smooth comparison test. For the later time period, we observed a decrease in the goal-looks for assertive–assertive condition.

4.2. Recall

Our aim was to analyze the proportion of correctly performed actions in response to different instructions. A generalized linear mixed effects model was fitted for each action type using glmer function in R with a binomial family (Bates, Mächler, Bolker, & Walker, 2015). The model included the factor instructions, coded using treatment-contrast ensuring that the intercept represented no-instruction condition (baseline) against which all instruction conditions were compared. Additionally, the model incorporated a full-random effect structure for the items and the subjects to account for potential variation arising due to complexity of items and individual differences. The random effects were added stepwise in the model following log-likelihood test to determine whether their inclusion significantly improved the model fit. The model revealed a between subjects ($SD = 0.42$) and items variance ($SD = 0.16$).⁴

A post-hoc individual pairwise comparison was conducted using the emmeans package to assess the main effect of instruction for each of the actions (Lenth, 2017). The model summary is provided in Table 2, and the results are visualized in Fig. 4. The model revealed that assertive–assertive instruction significantly improved recall for both kinds of actions as compared to the no-instruction (baseline) condition, suggesting that when actions are com-

Table 2

Experiment 1: Parameter estimates for the recall for contrastive and noncontrastive actions for levels of verbal instructions

Noncontrastive action	Estimate	SE	95% CI	<i>p</i>
Intercept	1.41	.29	[.83, 1.98]	< .001
Assertive–Assertive	.49	.18	[.14, .84]	.01
Assertive–Negative	−.17	.16	[−.49, .15]	.29
Negative–Assertive	.07	.17	[−.26, .40]	.67
Negative–Negative	−.08	.16	[−.40, .24]	.61
Contrastive action	Estimate	SE	95% CI	<i>p</i>
Intercept	1.28	.38	[.54, 2.02]	< .001
Assertive–Assertive	.78	.18	[.43, 1.14]	< .001
Assertive–Negative	.64	.18	[.29, .98]	< .001
Negative–Assertive	.42	.17	[.09, .75]	.01
Negative–Negative	.07	.16	[−.25, .39]	.68

Abbreviations: CI, confidence interval; Intercept, no-instruction condition.

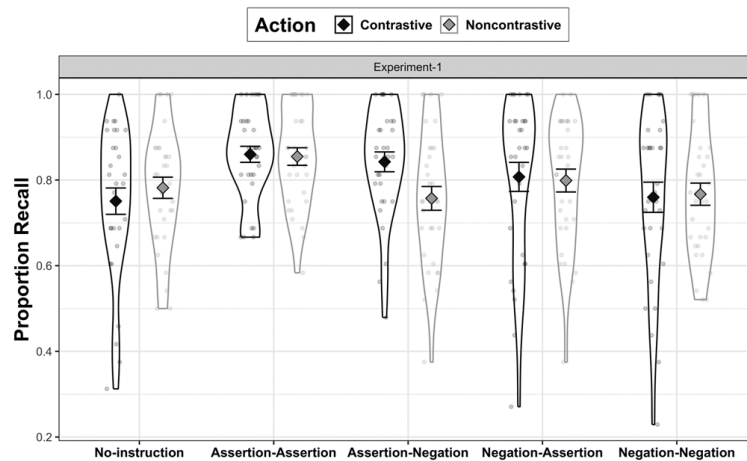


Fig. 4. Experiment 1: Proportion of correct recall for contrastive and noncontrastive actions. Instruction conditions are plotted against x-axis.

binéd with assertive instruction, remembering the sequence is enhanced as shown by the positive estimates for noncontrastive ($\beta = .49$, 95% CI [.14, .84], $p = .01$) and contrastive ($\beta = .78$, 95% CI [.43, 1.14], $p < .001$) actions. Crucially, the recall for the contrastive action also improved significantly from the baseline in the presence of the assertive–negative (contrastive) instruction ($\beta = .64$, 95% CI [.29, .98], $p < .001$). A similar pattern was also observed for the negative–assertive condition, another form of contrast, which also showed a positive effect on recall for contrastive actions ($\beta = .42$, 95% CI [.09, .75], $p = .01$). In summary, whereas participants showed an overall better recall for both types of action in the presence of assertive instructions, the assertive–negative and negative–assertive (contrastive instructions) specifically facilitated recall for the contrastive actions. Following this

result, we tested whether assertive–negative was more effective than only assertive instructions specifically for contrastive actions by doing a pairwise comparison test. The test revealed no significant difference, which indicates that the recall of a contrastive action under the assertive–negative (contrastive) was as good under the assertive–assertive instruction. For noncontrastive actions, only assertive instructions significantly improved action recall, suggesting that participants were more likely to score better on memory task when accompanied by only assertive (noncontrastive) instructions and that using negation for such actions was perhaps implausible. In contrast, for contrastive actions, recall was enhanced when paired with either assertive–negative (contrastive) or assertive–assertive (noncontrastive) instruction.

5. Discussion: Experiment 1

In our first experiment, we investigated the effect of verbal guidance on the attention to the goal and action recall. The verbal guidance consisted of verbal instructions, for which we used requests containing negation versus assertion, and combined them with the action demonstration for the first time. In this setup, the instructions were presented before the actions, and the offset of the instruction marked the beginning of an action (Fig. 1a). Our primary focus was on understanding how the time-dependent changes in the attention toward the goal shaped the action memory. The findings suggest that verbal instructions play a pivotal role in modulating the attention on the action goal. This is shown by a notable reduction in attention to the goal when participants received an assertive–negative instruction before a contrastive action.

In a contrastive action, the positive and its negative counterpart mapped onto a single semantics, which might have potentially brought the two paths following each other into a unified frame. In our experiment, the latter path was the alternate of the prior path, thus, the use of negation had two roles to play at the goal final object: inhibit the current subaction and highlight the absence of the prior subaction. The process of inhibition and reference to the previous subaction together might have contributed to the reduction in the goal-looks at the later time. Although guiding a contrastive action with a negation first (e.g., NOT Under–Now Under) also reduced the attention to the goal to some extent (see Fig. 3, top), this attenuation is higher when the negation is used for the later subaction (Now Over–NOT Over). This highlights the benefits of a prior context (assertive) before the negation can be used (Albu et al., 2021; Kaup et al., 2006; Wason, 1965).

For the noncontrastive action, reduction in the attention to the goal was noticeable for the assertive guidance at later time interval. In this case, the presence of an assertive guidance during the later subaction served to reiterate the preceding subaction.

Importantly, the modulation of attention from the goal affected the action memory. For contrastive context, we found that the memory was significantly enhanced when a negation was used at the later subaction (e.g., Now Over–NOT Over). This phenomenon can be attributed to the contextual processing of negation (Wason, 1965), wherein negation appears to be more informative by integrating prior and current subevents within a unified conceptual framework.

Crucially, according to the two-steps processing account, it has been suggested that when the positive counterpart is already under consideration, the activation of the alternate does not necessarily occur upon encountering negation (Kaup et al., 2007; Kaup & Dudschig, 2020). This alleviates processing costs and facilitates negation processing. Presenting the assertive instruction at the earlier subaction and negating the same action later—particularly in a context where the negated current subaction was previously affirmed—may have resulted in a single-step processing.

Additionally, we observed that assertive instructions enhanced the memory for both types of actions in general, which aligns with the previous findings that a multimodal presentation of an action enhances its perception (Brand & Tapscott, 2007; Hirsh-Pasek & Golinkoff, 1996; Rohlfing, Fritsch, Wrede, & Jungmann, 2006; Sciutti et al., 2016; Wrede, Schillingmann, & Rohlfing, 2013). Importantly, this perception has also been shown to be sensitive of how the modalities concerned are aligned (Bahrick & Lickliter, 2000, 2002; Bahrick et al., 2004). Although for a contrastive action, we observed a more reduction in the attention to the goal and better action memory in the presence of an assertive–negative instruction, this condition did not differ from the simple assertive instruction. One possible explanation is that not all participants might have benefited equally from negation, and individual differences in processing negation could have weakened the overall effect, especially given that negation is generally more cognitively demanding than its affirmative counterpart (Clark & Chase, 1972; Kaup & Dudschig, 2020).

The way in which verbal instructions can meddle with the action demonstration can be complex (Wolff & Holmes, 2011). Whereas in the first experiment, we explored the possibility that verbal instructions presented before actions could spotlight the upcoming action, another possibility is to offer language simultaneously to better match it with the ongoing actions. Since action perception is shown to be sensitive to the alignment of the modalities, whether sequential (Nomikou & Rohlfing, 2011; Sciutti et al., 2016), or simultaneous (Bahrick & Lickliter, 2000, 2002), we addressed the role of multimodal synchrony in a follow-up experiment. Particularly, in the second experiment, we aimed to investigate what consequence a different form of instruction-action synchrony has on the ever-present goal bias and action perception, such as when the instruction overlaps with the action.

6. Experiment 2: Instruction along with the action

In our follow-up experiment, we extended the approach of first experiment by introducing the verbal instruction along with actions, such that the instructions overlapped with the actions (Fig. 1b). In a study with infants, it was found that the timing of verbal instruction plays a crucial role in highlighting the subactions. For example, mothers introduced the verbal instruction before a subaction to direct attention and highlight it more than other components (Nomikou & Rohlfing, 2011; Sciutti et al., 2016). Conversely, studies also suggest that verbal instructions overlapping with actions mark distinct boundaries of an action, thus, facilitate the segmentation—a phenomenon known as “acoustic packaging” (Brand & Tapscott, 2007; Hirsh-Pasek & Golinkoff, 1996; Rohlfing et al., 2006; Wrede et al., 2013). It has been argued

that acoustic packaging essentially organizes a perception of meaningful units of actions by chunking them into subactions (Brand & Tapscott, 2007; Rohlfing et al., 2006; Schillingmann, Wrede, & Rohlfing, 2009; Wrede et al., 2013). Furthermore, presenting instructions overlapping with the subaction leads to a unitary perception of the modalities due to intersensory redundancy, a principal for which our cognition is developmentally shaped and generally offers an advantage for early learning (Bahrick & Lickliter, 2000, 2002; Bahrick et al., 2004). For such unitary patterns, the default behavior of our perceptual system is to prioritize the action goal, a tendency that also appears to be socially learned for most events due to sensorimotor processes (Lakusta & Landau, 2012; Wolff & Holmes, 2011). Thus, due to the instruction-action overlap in Experiment 2, we expected to observe greater attention to the action goal compared to Experiment 1.

7. Method

7.1. Participants

Forty-two university students (mean age = 24.45, $SD = 3.2$) participated in the experiment. All the participants were native or fluent German speakers. An informed consent was obtained from all the participants prior to data collection following the university ethical guidelines. Data from five participant were excluded from the analysis due to the track-loss for more than 50% of the trial or for failing to properly follow the instruction. Thus, the final sample size for the experiment was 37, which was comparable to the first experiment.

7.2. Stimuli

The stimuli were the same as those used in the first experiment, except that the verbal instructions now overlapped with the actions.

7.3. Procedure

Participants followed the same procedure and were briefed about the experiment in the same way as in the first experiment. Similarly, all the experiment sessions started with 10 practice trials, where participants familiarize themselves with the experiment setup and the objects.

7.4. Data processing

For all our analyses, we followed the procedure similar to our first experiment. This included using the same statistical models and coding schema to retain the uniformity across two experiments.

7.4.1. Eye movements preprocessing

Similar to the first experiment, we created a polygon AOI around the goal object. To retain the consistency of the AOI, we imported the AOI presets from our first experiment.

7.4.2. Preprocessing of recall videos

The recall videos from video cameras were compared with the videos shown on the eye tracker screen and coded in binary, 1 for correct and 0 for incorrect recall. The coding was performed by two student assistants providing the inter-rater reliability (Kendall's $W = 0.84$). Following this, we calculated the total error percentage and its distribution across error types. The coding yielded a total of 720 data points, with 130 classified as errors, resulting in an accuracy rate of 81.94%. Among the error types, omission errors were the most frequent (53.08%), followed by other errors (26.92%) and substitution errors (20%).⁵

8. Results

8.1. Eye movements

Following the analysis from our first experiment, we used GAM to fit the fixation data since the fixation pattern followed a nonlinear trend. The response variable was empirical-logit transformed prior to fitting the model following Barr (2008). We were specifically interested in examining the temporal pattern of attention directed toward the goal throughout the trial. A separate GAM model was fitted for each action type by incorporating the instruction as the parametric component and its factor smooth interaction by time. The predictor variables were treatment contrast-coded such that the model intercept represented the baseline condition (no-instruction). To select the best-fitting model, a log-likelihood model comparison was done between a reduced and full model following Barr et al. (2013). The Chi-Square test on the Maximum Likelihood (ML) scores indicated that model with instruction-time interaction was significantly better than the null model without interaction term, for contrastive ($\chi^2(8.00) = 751.35, p < .001$) and noncontrastive action ($\chi^2(8.00) = 648.48, p < .001$). The model summary is provided in Table 3, and the estimated effects are visualized in Fig. 5, respectively.

(A) Contrastive action: Similar to the first experiment, when instruction conditions were compared to the baseline (no-instruction), the parametric terms (time-independent) revealed a significant effect for the instructions, such that attention to the goal significantly decreased for assertive-negative condition ($\beta = -.037, p < .001$). Crucially, the assertive-negative instruction was contrastive, replicating our previous results from the first experiment. Moreover, the effects observed were more pronounced than in the previous experiment (Fig. 5, top). In contrast, other instruction conditions, such as assertive-assertive and negative-negative, led to a significant increase in goal-looks from the baseline, as shown by the positive estimates for assertive-assertive ($\beta = .016, p < .001$) and negative-negative ($\beta = .019, p < .001$) instruction conditions.

Following this, we examined the time-dependent terms and found that all smooth terms were significantly different from the reference level, indicating distinct temporal trends across conditions. Notably, as in the first experiment, the assertive-negative condition showed the lowest effective degrees of freedom (edf = 8.76, $p < .001$), with its effect emerging earlier than in Experiment 1. To further investigate this, we additionally conducted a GAM difference smooth comparison for each instruction condition against the baseline (no-instruction). Consistent with the first experiment, the difference plot revealed that assertive-

Table 3

Experiment 2: Estimates of looks toward the goal for contrastive and noncontrastive actions

Contrastive action			
Parametric coefficients	Estimate	SE	<i>p</i>
Intercept	−3.75	.003	< .001
Assertive–Assertive	.016	.004	< .001
Assertive–Negative	−.037	.004	< .001
Negative–Assertive	−.006	.004	.18
Negative–Negative	.019	.004	< .001
Smooth terms	edf	<i>F</i>	<i>p</i>
s(bin):No-instruction	8.98	364.68	< .001
s(bin):Assertive–Assertive	8.97	475.49	< .001
s(bin):Assertive–Negative	8.76	215.45	< .001
s(bin):Negative–Assertive	8.95	280.78	< .001
s(bin):Negative–Negative	8.96	364.45	< .001
Noncontrastive action			
Parametric coefficients	Estimate	SE	<i>p</i>
Intercept	−3.74	0.003	< .001
Assertive–Assertive	−.013	.004	< .001
Assertive–Negative	.003	.004	.46
Negative–Assertive	−.034	.004	< .001
Negative–Negative	−.024	.004	< .001
Smooth terms	edf	<i>F</i>	<i>p</i>
s(bin):No-instruction	8.96	204.13	< .001
s(bin):Assertive–Assertive	8.95	172.11	< .001
s(bin):Assertive–Negative	8.91	77.72	< .001
s(bin):Negative–Assertive	8.86	95.91	< .001
s(bin):Negative–Negative	8.95	151.08	< .001

Abbreviation: Intercept, no-instruction condition.

negative condition led to a reduction in goal-looks, followed by the assertive–assertive condition.⁶

(B) Noncontrastive action: For the parametric (time-independent) component, there was a significant difference in the goal-looks between baseline (no-instruction) and other instruction conditions, such that fixation to goal significantly decreased for all the instruction conditions except when assertive–negative instruction was provided ($\beta = .003$, $p = .46$). For the time-dependent smooth terms, we observed significant different temporal pattern for all the instruction conditions as compared to the reference level. Although we did not find a lower edf score for the assertive–assertive instruction, as was observed in Experiment 1, a GAM difference smooth comparison revealed that fixation on the goal decreased significantly in both the assertive–assertive and assertive–negative conditions compared to the baseline.

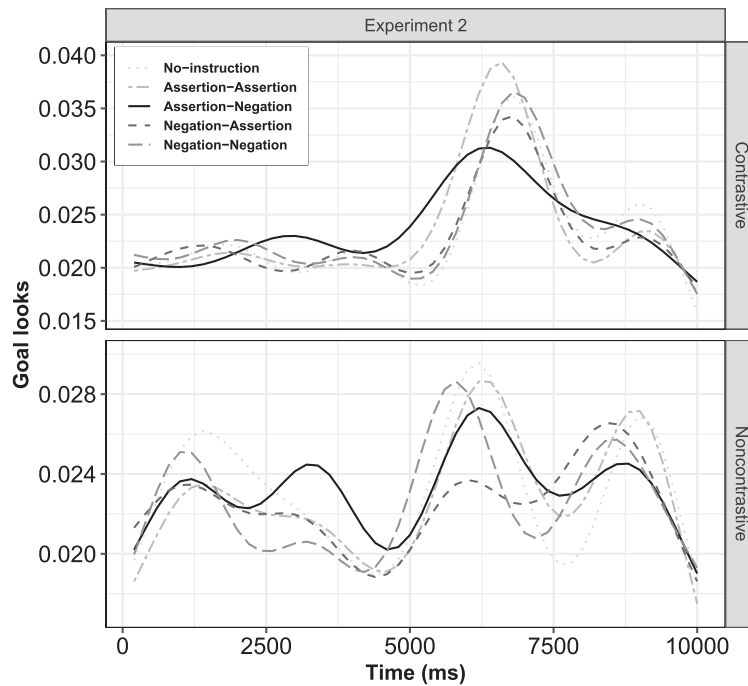


Fig. 5. Experiment 2: Estimated goal-looks for noncontrastive action (top plot) and contrastive action (bottom plot).

Overall, our finding for the contrastive action replicated our results from the first experiment. We found reduced attention to the goal, compared to the baseline, when a contrastive action was paired with an assertive–negative (contrastive) instruction, as shown by the smooth difference comparisons. For the noncontrastive actions, in Experiment 2, we observed a decrease in goal-looks for both assertive–assertive and assertive–negative conditions compared to the baseline (no-instruction). While the effect for the assertive–assertive condition replicated the findings from the first experiment, the decrease in attention to the goal in the assertive–negative condition in the second experiment was unexpected and might reflect the benefit of using an assertion at the first subevent for noncontrastive actions. Notably, the peaks in the goal-looks occurred earlier in this experiment than in the first experiment, and the effect was more pronounced. While the influence of multimodal overlapping on attention was already visible here, turning back to our hypothesis on instruction-action temporal synchrony, we analyzed early and late goal-look behaviors separately by statistically comparing the two experiments (see Section 9.3).

8.2. Recall

For the analysis of the recall data, we used our earlier generalized linear mixed model to fit the complete dataset across both experiments, and conducted a log-likelihood test between the full and reduced models (Barr et al., 2013). When compared across experiments, we did not observe a significant increase in the score for the second experiment (instruction along with the action), suggesting that the recall remains invariant to the instruction-action synchrony.

Table 4
Experiment 2: Parameter estimates for the recall for contrastive and noncontrastive actions for levels of verbal instructions

Noncontrastive action	Estimate	SE	95% CI	<i>p</i>
Intercept	1.33	.15	[1.03, 1.63]	< .001
Assertive–Assertive	.63	.16	[-.32, .95]	< .001
Assertive–Negative	–.15	.14	[–.43, .14]	.31
Negative–Assertive	.33	.15	[.03, .62]	.03
Negative–Negative	.00	.15	[–.29, .29]	1.00
Contrastive action	Estimate	SE	95% CI	<i>p</i>
Intercept	1.64	.20	[1.24, 2.04]	< .001
Assertive–Assertive	.36	.20	[–.03, .76]	.07
Assertive–Negative	.55	.17	[.22, .89]	< .001
Negative–Assertive	–.16	.19	[–.52, .20]	.39
Negative–Negative	–.17	.15	[–.47, .13]	.28

Abbreviations: CI, confidence interval; Intercept, no-instruction condition.

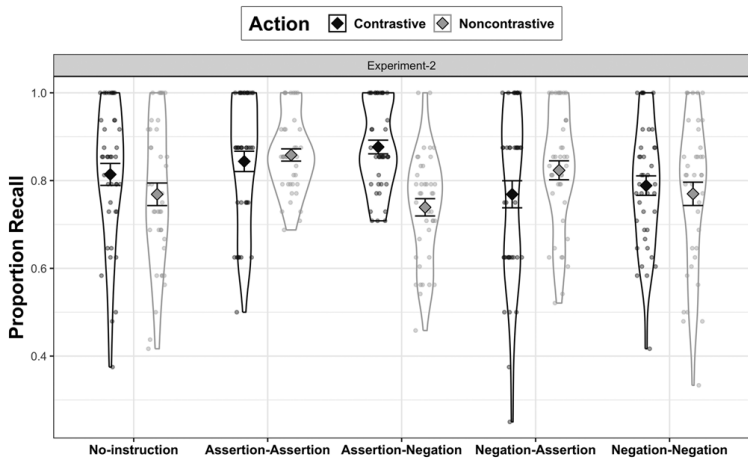


Fig. 6. Experiment 2: Proportion of correct recall for contrastive and noncontrastive actions. Instruction conditions are plotted against x-axis.

Following this, we tested the main effect of instruction on action recall focusing specifically on Experiment 2. The model included the main predictors by treatment-coding, such that the intercept represented the no-instruction (baseline) condition. To account for the variances arising due to the items or the subjects, both were treated as the random factors. Adding the by-item adjustment to the intercept, however, did not significantly improve the model. A post-hoc multiple comparison was conducted using the emmeans package to obtain the estimates for the factors of interest after the model fitting. The model summary is provided in Table 4, and the estimates are visualized in Fig. 6.

Beginning with the effect of assertive instruction when compared against baseline, we found that an assertive–assertive instruction led to a significant increase in recall only for the

noncontrastive action ($\beta = .63$, 95% CI [.32, .95], $p < .001$) but not for the contrastive action ($\beta = .36$, 95% CI [−.03, .76], $p = .07$). This pattern was different from the first experiment (where the instruction preceded the action), in which assertive instruction facilitated memory for both types of actions. Conversely, in Experiment 2, we observed that an assertive–assertive (noncontrastive) instruction particularly facilitating recall for the noncontrastive action (see Table 4). For the contrastive instruction, we observed that assertive–negative instruction significantly improved the recall for the contrastive action, as indicated by a positive estimate ($\beta = .55$, CI [.22, .89], $p < .001$). This result was consistent with the findings from the first experiment, where assertive–negative (contrastive) instructions were shown to be beneficial only for contrastive but not for noncontrastive action. Following this, we tested whether a different form of verbal contrast (negative–assertive) significantly increased the memory for contrastive action as observed in the first experiment. Contrary to the first experiment, in the second experiment, a negative–assertive instruction did not facilitate memory for the contrastive actions. To test whether assertive–negative was significantly better than only assertive instructions for contrastive action, a post-hoc pairwise comparison was applied using emmeans package (Lenth, 2017). The comparison revealed no significant difference between the assertive–assertive and assertive–negative conditions, suggesting that although recall for contrastive actions was higher under assertive–negative instructions, the improvement was not statistically significant compared to assertive-only guidance. In summary, when compared to the first experiment, two findings are particularly notable: (a) a verbal contrast in the form of assertive–negative played a crucial role in particularly guiding the contrastive actions, and (b) assertive–assertive instructions selectively enhanced memory for the noncontrastive action—a finding which diverged from the first experiment, where assertive–assertive facilitated memory for both types of actions, irrespective of whether it was contrastive or noncontrastive.

8.3. *Effect of instruction–action synchrony on attention to the goal*

To evaluate the effect of instruction synchrony on the attention to the goal, we analyzed the data across two experiments. Our predictions were derived from two rationales: Whereas in Experiment 1, we explored the possibility that presenting the instruction as a spotlight before the subaction might interfere with early sensorimotor processes, in Experiment 2, we expected that the overlapping presentation of the modalities would lead to a unitary perception—thereby naturally directing attention toward the action goal as a result of automatized sensorimotor routines (Bahrick & Lickliter, 2000; Bahrick et al., 2004; Zacks & Swallow, 2007). Accordingly, we predicted that attention to the action goal would be higher in Experiment 2 compared to Experiment 1.

For the analysis, we only included the trials, where verbal instructions were provided by excluding the baseline (no-instruction) trials, since the effect of synchrony would only be present when there is an instruction. The data was fitted on the previous GAM model by including experiment number as the main predictor. To observe a temporal difference in attention to the goal across two forms of synchrony, a time-dependent test of significance was done using mgcv package in R (Wood, 2011). The overall difference in the attention toward the goal across two experiments is shown in Fig. 7, and the estimates of comparisons

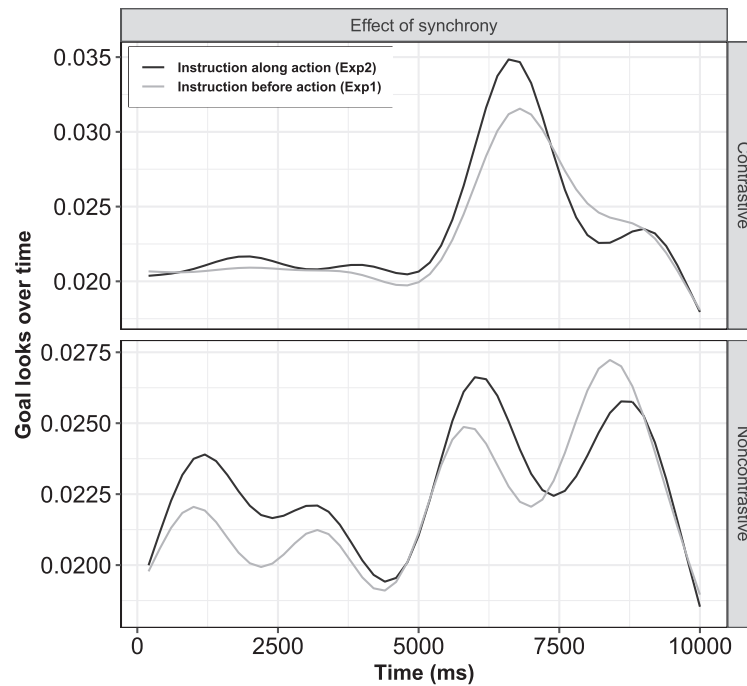


Fig. 7. Effect of forms of synchrony on goal-looks over time for contrastive and noncontrastive actions. The error-band represents 95% CI.

Table 5

Comparison between first and second experiment for the effect of synchrony on attention to the goal

Contrastive action			
Parametric coefficients	Estimate	SE	<i>p</i>
Intercept	−3.76	.01	< .001
Instruction along action	.02	.002	< .001
Smooth terms	edf	<i>F</i>	<i>p</i>
s(bin):Instruction before action	8.95	811.40	< .001
s(bin):Instruction along action	8.99	1246.30	< .001
Noncontrastive action			
Parametric coefficients	Estimate	SE	<i>p</i>
Intercept	−3.79	.001	< .001
Instruction along action	.03	.002	< .001
Smooth terms	edf	<i>F</i>	<i>p</i>
s(bin):Instruction before action	8.97	435.30	< .001
s(bin):Instruction along action	8.98	438.10	< .001

are shown in Table 5. The parametric terms suggested that when instruction overlapped with the action (instruction along action), it resulted in significantly higher looks to the goal, as shown by positive estimates for both contrastive ($\beta = .02$, $p < .001$) and noncontrastive actions ($\beta = .03$, $p < .001$). Conversely, presenting instructions before the action reduced

Table 6
Parameter estimates for the recall score and fixation

Terms	Estimate	SE	95% CI	p
Intercept	−4.87	.23	[−5.32, 4.42]	< .001
c_mean_recall	−.26	.04	[−.32, −.17]	< .001

attention to the goal (Fig. 7). We also observed that the effect of synchrony on looks to the goal emerged earlier for the noncontrastive than contrastive actions. Given the limited studies investigating oculomotor behavior in contrastive motion events, we refrained from making a concrete prediction on this aspect. However, previous research has suggested that path complexity influences eye movements, such that complex paths are associated with greater cognitive load, leading to more frequent regressions and increased fixation durations (Kosch, Hassib, Woźniak, Buschek, & Alt, 2018). Based on this, we suggest that the earlier fixation on the goal in the noncontrastive condition might be due to the relative simplicity of the unidirectional paths, compared to the bidirectional nature of contrastive paths. That said, it remains unclear whether changes in direction were actually perceived as complex within the flow of continuous motion events, a question that remains open for future investigation.

8.4. Does attention to the goal influence action recall?

Whereas the multimodal synchrony affected the attention to the goal, we conducted an analysis to investigate whether it relates to the action memory. Specifically, we examined the relationship between the attention to the goal and action memory. The aim was to test whether the recall scores negatively predict the overall looks toward the goal. This will inform us that whether high attention to the goal leads to a coarse perception of intermediate events, ultimately shaping the overall memory of the action. Thus, the data from two experiments was modeled together by treating the recall score as the predictor and fixation on the action goal as an outcome variable. To preserve the time dimension, we followed a method previously used to analyze visual-world data (Porretta et al., 2018; Porretta et al., 2016; Wieling, 2018). The recall score was mean-centered for better convergence. To account for variability in looking patterns between subjects and items, a random intercept adjustment to the subject and item was added to the model. This decision was based on a “goodness of fit” test comparing models with and without random adjustments. The random effects structure captures differences such as whether certain participants are early or late in looking at the goal, and whether specific items influence distinct looking patterns. We expected that the recall score would negatively predict the fixation on the goal if there exists an inverse relation between the action memory and the attention to the goal. The model comparison was done using log-likelihood method, which indicated that including the recall score in the model led to a significant better fit ($\chi^2(8.00) = 1244.67, p < .001$).

The predictions are shown in Table 6, and the results are visualized in Fig. 8. There was a significant negative effect of score such that the recall score negatively predicted the looks toward the goal ($\beta = -.26$, 95% CI [−.32, −.17], $p < .001$). To put it in other words, participants’ recall was better when they managed to “de-bias” themselves from the goal. The

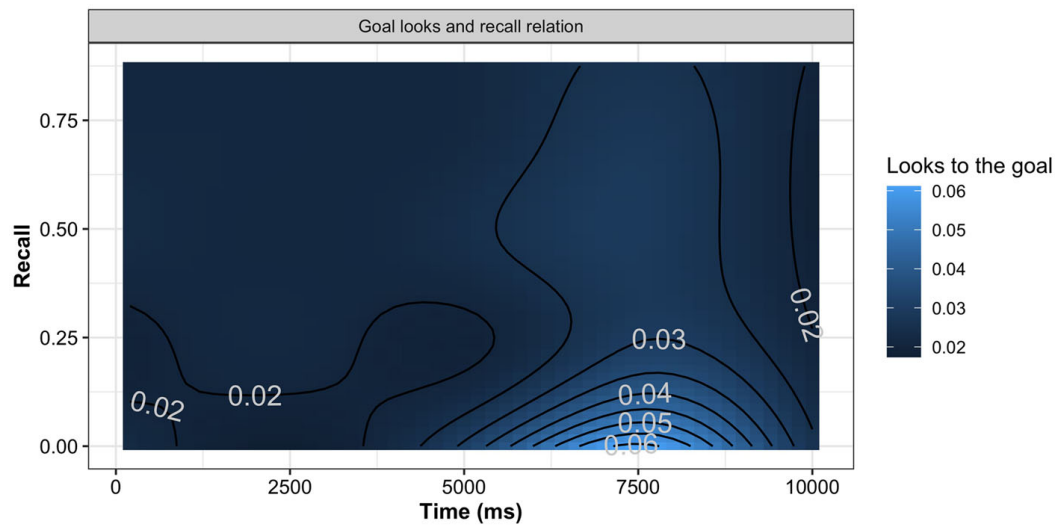


Fig. 8. Plot for the attention to the goal and action recall. Y-axis shows the recall score and x-axis shows the time from beginning of the trial. The numbers indicate the probability of looks toward the goal with white shading an increase in the goal-looks.

effect predominantly occurred approximately between 5000 and 9000 ms window peaking at approximately 7500 ms. This was also the peak time for the looks toward the goal across two experiments as observed for different forms of modalities alignment (Fig. 7). Additionally, the model revealed a variation between subjects ($SD = 1.40$) and items ($SD = 0.31$).⁷

9. Discussion: Experiment 2

The aim of the second experiment was to address other possibilities of language interplaying with the actions. In the first experiment, the instruction was presented before the action, in the second experiment, we presented the instruction along with the action to “package” it by overlapping the modalities (Brand & Tapscott, 2007; Hirsh-Pasek & Golinkoff, 1996; Schillingmann et al., 2009). We introduced the overlapping to validate our findings for the goal attention and action memory across two forms of instruction–action synchrony. With overlapping modalities, we observed that the verbal contrast (assertive–negative) led to a more pronounced reduction in the attention to the goal for the contrastive action. The co-occurrence of the instruction and action has been proposed to enhance one’s ability to see the sequence of actions as unitary, a phenomenon rooted in developmental processes and considered relevant for action learning (Bahrick et al., 2004; Brand & Tapscott, 2007). It is also worth noting that interpreting negations has always been difficult than assertions (Carpenter & Just, 1975; Clark & Chase, 1972), also in the context of actions (de Vega et al., 2016; Tettamanti et al., 2008). Hence, overlapping the negation with an action might have made the path of an action easier to interpret as opposed to when negation is presented before the action. Thus, the verbal guidance comprising negation is more sensitive to the multimodal overlap as compared to an assertive guidance. This was particularly observable for the recall. In the second exper-

iment, we observed a recall pattern similar to that of the first experiment, with two notable differences: (a) Whereas in the first experiment, assertive instructions facilitated recall for both contrastive and noncontrastive actions, in the second experiment, assertive instructions selectively enhanced memory for noncontrastive actions, and similarly, (b) whereas recall for contrastive actions in the first experiment was facilitated by both forms of verbal contrast (assertive–negative or negative–assertive), in the second experiment, only assertive–negative instruction significantly improved recall for contrastive actions. Thus, compared to Experiment 1, in Experiment 2, when the instruction overlapped with the action, the role of verbal instruction became more tailored to the specific type of action it guided.

10. General discussion

To enhance the demonstration of actions, verbal instructions are used. They can contain assertive or negative information. Whereas many studies reveal the effect of assertion versus negation processing, an embeddedness of this processing in the context of an action demonstration is lacking but could be insightful for the setting of joint work and learning.

For our study, we linked together research from action demonstration with linguistic research on negation to offer effective verbal guidance. We conducted two eye tracking experiments where the verbal instructions were presented either before the action (Experiment 1) or along with the action (Experiment 2). This was done to foster two forms of synchrony between instruction and action, in line with the theory of “acoustic packaging” (Brand & Tapscott, 2007, p. 323) and intersensory redundancy (Bahrick et al., 2004). Crucially, the instructions were a sequence of assertion and negation, combined with action presentation that was either contrastive or noncontrastive. To remind the reader, with a contrastive action, we refer to two subactions that contrast with each other (e.g., OVER–UNDER), while with noncontrastive action, we refer to two recurrent subactions (e.g., OVER–OVER).

Our findings shed light on how the timing and the nature of instruction influence the attention and action memory. We observed that contrastive instructions (assertive–negative) attenuated the attention from the goal for contrastive actions. This pattern was visible in both forms of instruction–action synchrony, regardless of whether the instructions were presented before or along with the action.

When comparing across the experiments, we found that overlapping the modalities led to an increased attention on the goal, compared to when the instructions were presented right before the action (Fig. 7). This was in line with our predictions regarding instruction–action synchrony (Fig. 1), which was grounded in the Intersensory Redundancy Hypothesis (IRH) (Bahrick & Lickliter, 2000, 2002; Bahrick et al., 2004). According to IRH, the co-occurrence of modalities leads to a reduced selective attention on any one modality, resulting in a unitary perception (Bahrick et al., 2004). This unitary perception plays an ecological role in early learning by supporting the integration of multimodal properties, which lays the foundation for perceptual processing in adulthood. Since these properties are foundational and sometimes considered to be a part of sensorimotor routines (Zacks et al., 2001), the default behavior of our perceptual system is to prioritize action goal in such situations, unless

attention is ostensibly directed toward a specific subaction, through language spotlighting (Wolff & Holmes, 2011). This was the case when instructions were presented before the action. In this context, the instructions rather served to ostensibly introduce and highlight a particular subaction to which it was intended to guide. Thus, presenting instruction before the subaction might ensure sustained attention to that particular subaction due to selective attention on each modality at a given time (here: verbal or visual). Given that there has been a trade-off between goal-attention and the intermediary components (Ikegami, 1979; Papafragou, 2010; Zacks & Swallow, 2007), the overlapping condition in Experiment 2 may have resulted in reduced selective attention to either modality, manifesting as an inherent tendency, such as increased looks toward the action goal. Thus, the action–instruction synchrony involves a delicate balance between facilitating attention to the goal and subactions, which might serve different purposes for communication. To highlight an immediate subaction, providing an instruction before the action can enhance sustained attention on that particular subaction. Conversely, if the purpose is to ensure more attention on the goal than the mid-events, presenting instructions along with the action can be more beneficial. This is the case for children, who certainly need to discern the relevance of the action for a specific goal, and thus, it is not surprising that action–verbal instruction synchrony was found to be high with younger children (Nomikou & Rohlfing, 2011). For event cognition research, it is important to emphasize that experiments investigating online attention to the goal during event conceptualization should carefully take the timing of instruction in consideration. This is particularly crucial, as attention to the goal appears to be sensitive to instruction–action synchrony.

Our study further highlights that the contrastive instruction differs from mere assertive instruction due to the role of negation in generating a rich contextual effect that is appropriate for a contrastive action. In a contrastive action, the later subaction was the opposite of the prior subaction. Thus, negating the later subaction entails referring to the previous subaction. This unfolding context at verbal and visual level might have led to an easier interpretation of the whole action. The interpretation aligns with the premise that contrastiveness relies on connecting later subevent with the preceding ones (Talmy, 1985, 1975, 2000), and when the modalities temporally overlap, this connection is more profound (Brand & Tapscott, 2007; Gogate, Bahrick, & Watson, 2000; Hirsh-Pasek & Golinkoff, 1996; Rohlfing et al., 2006). Furthermore, this also connects with findings in the literature showing that people often use negation to prevent the repetition of a past action and to inhibit a prepotent or biased response (e.g., Beltrán et al., 2021; de Vega et al., 2016). Our results correspond with this, suggesting that contrastive instructions may serve a similar inhibitory function by modulating attention and facilitating memory for contrastive actions. For example, in a teaching scenario, instructors often use negation to achieve these dual objectives—preventing repetitive actions and inhibiting bias responses—due to negation’s ability to influence domain-general inhibitory mechanisms. As an extension, an effective approach for guiding an action should employ a forward referencing strategy in addition to a negation. In this approach, one not only inhibits the response bias by referring back and contrasting with past action(s), but also states what needs to be done in place, as in the example, “Do not move it right, but left!” A recent study has utilized this approach to show that contrasting

two opposing manners in an action fosters learning of a targeted manner (Groß et al., 2023). In this context, negation can serve as a powerful linguistic tool for guiding such behaviors by inhibiting expected responses and contrasting them with the past action(s). Crucially, contrastive guidance strategy has the potential to attenuate the attention directed toward the goal that arises from repeated expectations. It would also be intriguing to explore the broader implications of such guidance in real-world scenarios, investigating whether these effects extend beyond actions and impact memory retention over time. This exploration could have significant implications for learning or understanding actions, particularly in terms of how individuals process and retain attenuated information, such as the manner of an action, in addition to the salient properties, such as path (Papafragou et al., 2008; Papafragou, 2010). In this regard, our findings revealed that attenuated attention from the goal led to a more holistic understanding of an action. This finding suggests a fundamental trade-off between attention to the goal and event memory: when attention is biased toward the goal, other preceding subevents might be encoded less robustly. This resonates with the previous findings, such as the source–goal asymmetry in event memory (Lakusta & Landau, 2012; Papafragou, 2010), encoding of the path and the manner (Papafragou, 2010; Skordos et al., 2020), and prioritization of the goal over the means (Bird, Brindley, Leighton, & Heyes, 2007). Our study also highlights the potential use of “acoustic packaging” in such scenarios. By packaging the intermediate subactions and highlighting the fine-grained aspects, acoustic packaging may mitigate the detrimental effects of the potential biases on event memory. Our study was limited by examining the attention particularly on the goal and we did not directly measure the attention on the subactions. Future studies should fill this gap by simultaneously assessing the attention on the subaction(s) and the action goal to observe the real-time attentional trade-off between these two constituents in an action, particularly for two forms of multimodal synchrony.

In conclusion, our study demonstrated the robustness of assertive instructions in enhancing the action recall and highlighted the potential of negation to guide contrastive actions. It further suggests a delicate balance between the goal attention and event memory underlying the action perception. In our study, participants’ recall was better when they managed to “de-bias” themselves from the goal. Future studies should delve into the question of whether and how visual and verbal contrasts together shape our memory and everyday perception, particularly when used in an interactional context Supporting Information.

Acknowledgments

This research is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of project A05 TRR 318/1 2021 “Constructing Explainability” - 43844582. We thank the reviewers for their constructive comments, which greatly contributed to the improvement of this manuscript. We also acknowledge the valuable support of our student assistants, Alana Vandekerkhof, Michelle Gollai, and Annabell Horch, in data collection and annotation.

Open access funding enabled and organized by Projekt DEAL.

Data availability statement

The data and analysis codes that support the findings of this study are openly available in OSF at https://osf.io/zq7b2/?view_only=f571b0c6f306476792f20e963db3ca4c

Notes

- 1 Information on sample size calculation is provided in related OSF repository.
- 2 A table of the error types is provided in the Supplementary Materials.
- 3 GAM difference plots are provided in the Supplementary Materials.
- 4 The random effects table and plots are provided in the Supplementary Materials.
- 5 A full table of each error type is provided in the Supplementary Materials.
- 6 GAM difference plots are provided in the Supplementary Materials.
- 7 Estimates and plots for random effect structure are provided in the Supplementary Materials.

References

- Albu, E., Tsaregorodtseva, O., & Kaup, B. (2021). Contrary to expectations: Does context influence the processing cost associated with negation? *Journal of Psycholinguistic Research*, 50(6), 1215–1242.
- Altmann, G. T. M., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 347–386). Routledge.
- Ambrosini, E., Reddy, V., de Looper, A., Costantini, M., Lopez, B., & Sinigaglia, C.. (2013). Looking ahead: Anticipatory gaze and motor ability in infancy. *PLOS ONE*, 8(7), e67916.
- Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, 36(2), 190–201.
- Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. *Advances in Child Development and Behavior*, 30, 153–187.
- Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science*, 13(3), 99–102.
- Banh, N. C., Tünnermann, J., Rohlfing, K. J., & Scharlau, I. (2024). Benefiting from binary negations? Verbal negations decrease visual attention and balance its distribution. *Frontiers in Psychology*, 15, 1451309.
- Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, 59(4), 457–474.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4), 577–609.
- Bartoli, E., Tettamanti, A., Farronato, P., Caporizzo, A., Moro, A., Gatti, R., Perani, D., & Tettamanti, M. (2013). The disembodiment effect of negation: Negating action-related sentences attenuates their interference on congruent upper limb movements. *Journal of Neurophysiology*, 109(7), 1782–1792.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Beltrán, D., Liu, B., & de Vega, M. (2021). Inhibitory mechanisms in the processing of negations: A neural reuse hypothesis. *Journal of Psycholinguistic Research*, 50(6), 1243–1260.
- Bird, G., Brindley, R., Leighton, J., & Heyes, C. (2007). General processes, rather than "goals," explain imitation errors. *Journal of Experimental Psychology: Human Perception and Performance*, 33(5), 1158–1169.

- Bloom, L. (1970). *Language development: Form and function in emerging grammars*. Cambridge, MA: MIT Press.
- Bloom, L. (1993). *Language development from two to three*. New York: Cambridge University Press.
- Brand, R. J., & Tapscott, S. (2007). Acoustic packaging of action sequences by infants. *Infancy*, 11(3), 321–332.
- Bunger, A., Skordos, D., Trueswell, J. C., & Papafragou, A. (2016). How children and adults encode causative events cross-linguistically: Implications for language production and attention. *Language, Cognition and Neuroscience*, 31(8), 1015–1037.
- Bunger, A., Skordos, D., Trueswell, J. C., & Papafragou, A. (2021). How children attend to events before speaking: Crosslinguistic evidence from the motion domain. *Glossa*, 6(1), Article 1210.
- Bunger, A., Trueswell, J. C., & Papafragou, A. (2012). The relation between event apprehension and utterance formulation in children: Evidence from linguistic omissions. *Cognition*, 122(2), 135–149.
- Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, 82(1), 45–73.
- Choi, S. (1988). The semantic development of negation: A cross-linguistic longitudinal study. *Journal of Child Language*, 15(3), 517–531.
- Clark, H. H., Carpenter, P. A., & Just, M. A. (1973). On the meeting of semantics and perception. In W. G. Chase (Ed.), *Visual information processing* (pp. xiv, 555–xiv, 555). Oxford: Academic Press.
- Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, 3(3), 472–517.
- Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, 13(4), 148–153.
- Dale, R., & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science*, 35(5), 983–996.
- de Vega, M., Morera, Y., León, I., Beltrán, D., Casado, P., & Martín-Loeches, M. (2016). Sentential negation might share neurophysiological mechanisms with action inhibition: Evidence from frontal theta rhythm. *Journal of Neuroscience*, 36(22), 6002–6010.
- Drozd, K. F. (1995). Child English pre-sentential negation as metalinguistic exclamatory sentence negation. *Journal of Child Language*, 22(3), 583–610.
- Dudschig, C., Kaup, B., Liu, M., & Schwab, J. (2021). The processing of negation and polarity: An overview. *Journal of Psycholinguistic Research*, 50(6), 1199–1213.
- Glenberg, A. M., & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex*, 48(7), 905–922.
- Glenberg, A. M., Robertson, D. A., Jansen, J. L., & Johnson-Glenberg, M. C. (1999). Not propositions. *Cognitive Systems Research*, 1(1), 19–33.
- Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, 71(4), 878–894.
- Gredebäck, G., & Falck-Ytter, T. (2015). Eye movements during action observation. *Perspectives on Psychological Science*, 10(5), 591–598.
- Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11(4), 274–279.
- Groß, A., Singh, A., Banh, N. C., Richter, B., Scharlau, I., Rohlfing, K. J., & Wrede, B. (2023). Scaffolding the human partner by contrastive guidance in an explanatory human–robot dialogue. *Frontiers in Robotics and AI*, 10, Article 1236184.
- Hespos, S. J., Grossman, S. R., & Saylor, M. M. (2010). Infants' ability to parse continuous actions: Further evidence. *Neural Networks*, 23(8), 1026–1032.
- Hespos, S. J., Saylor, M. M., & Grossman, S. R. (2009). Infants' ability to parse continuous actions. *Developmental Psychology*, 45(2), 575–585.
- Hirsh-Pasek, K., & Golinkoff, R. M. (1996). *The origins of grammar: Evidence from early language comprehension*. Cambridge, MA: MIT Press.
- Holmes, K. J., & Wolff, P. (2010). Simulation from schematics: Dorsal stream processing and the perception of implied motion. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32.

- Horn, L. R. (2001). *A natural history of negation*. The David Hume Series. Center for the Study of Language and Information.
- Huff, M., Papenmeier, F., & Zacks, J. M. (2012). Visual target detection is impaired at event boundaries. *Visual Cognition*, 20(7), 848–864.
- Hummer, P., Wimmer, H., & Antes, G. (1993). On the origins of denial negation. *Journal of Child Language*, 20(3), 607–618.
- Ikegami, Y. (1979). ‘Goal’ over ‘source’: A case of dissymmetry in linguistic orientation. *Angol Filológiai Tanulmányok / Hungarian Studies in English*, 12, 139–157.
- Kanakogi, Y., & Itakura, S. (2011). Developmental correspondence between action prediction and motor ability in early infancy. *Nature Communications*, 2, 341.
- Kaup, B., & Dudschig, C. (2020). Understanding negation: Issues in the processing of negation. In V. Déprez & M. T. Espinal (Eds.), *The Oxford handbook of negation* (pp. 635–655). Oxford University Press.
- 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(7), 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* (pp. 255–288). Lawrence Erlbaum Associates Publishers.
- Kosch, T., Hassib, M., Woźniak, P. W., Buschek, D., & Alt, F. (2018). Your eyes tell: Leveraging smooth pursuit for assessing cognitive workload. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, (pp. 1–13). New York: Association for Computing Machinery.
- Kosie, J. E., & Baldwin, D. A. (2021). Dwell times showcase how goal structure informs preschoolers’ analysis of unfolding motion patterns. *Child Development*, 92, 2235–2243.
- Lakusta, L., & Landau, B. (2005). Starting at the end: The importance of goals in spatial language. *Cognition*, 96(1), 1–33.
- Lakusta, L., & Landau, B. (2012). Language and memory for motion events: Origins of the asymmetry between source and goal paths. *Cognitive Science*, 36(3), 517–544.
- Lüdtke, J., & Kaup, B. (2006). Context effects when reading negative and affirmative sentences. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.
- Lenth, R. V. (2017). emmeans: Estimated marginal means, aka least-squares means. <https://cran.r-project.org/package=emmeans>.
- Levelt, W. J. (1993). *Speaking: From intention to articulation*. MIT Press.
- Mayo, R., Schul, Y., & Rosenthal, M. (2014). If you negate, you may forget: Negated repetitions impair memory compared with affirmative repetitions. *Journal of Experimental Psychology: General*, 143(4), 1541–1552.
- Monroy, C., Gerson, S., & Hunnius, S. (2017). Infants’ motor proficiency and statistical learning for actions. *Frontiers in Psychology*, 8, Article 2174.
- Nomikou, I., & Rohlfing, K. (2011). Language does something: Body action and language in maternal input to three-month-olds. *IEEE Transactions on Autonomous Mental Development*, 3(2), 113–128.
- Nordmeyer, A. E., & Frank, M. C. (2023). Pragmatic felicity facilitates the production and comprehension of negation. *Collabra: Psychology*, 9(1), 67931.
- Olofson, E. L., & Baldwin, D. (2011). Infants recognize similar goals across dissimilar actions involving object manipulation. *Cognition*, 118(2), 258–264.
- Orenes, I., Beltrán, D., & Santamaría, C. (2014). How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language*, 74, 36–45.
- Orenes, I., Moxey, L., Scheepers, C., & Santamaría, C. (2016). Negation in context: Evidence from the visual world paradigm. *Quarterly Journal of Experimental Psychology (2006)*, 69(6), 1082–1092.
- Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. *Cognitive Science*, 34(6), 1064–1092.
- Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition*, 108(1), 155–184.

- Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, 106(2), 665–681.
- Porretta, V., Kyröläinen, A.-J., Van Rij, J., & Järvikivi, J. (2018). Visual world paradigm data: From preprocessing to nonlinear time-course analysis. In *Intelligent Decision Technologies 2017*, volume 73 (pp. 268–277). Cham: Springer International Publishing.
- Porretta, V., Tucker, B. V., & Järvikivi, J. (2016). The influence of gradient foreign accentedness and listener experience on word recognition. *Journal of Phonetics*, 58, 1–21.
- Posit Team. (2024). Rstudio: Integrated development environment for R.
- R Core Team. (2022). R: A language and environment for statistical computing.
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169–192.
- Rohlfing, K., Fritsch, J., Wrede, B., & Jungmann, T. (2006). How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, 20(10), 1183–1199.
- Saslow, M. G. (1967). Latency for saccadic eye movement. *Journal of the Optical Society of America*, 57(8), 1030–1033.
- Schillingmann, L., Wrede, B., & Rohlfing, K. J. (2009). A computational model of acoustic packaging. *IEEE Transactions on Autonomous Mental Development*, 1(4), 226–237.
- Sciutti, A., Lohan, K. S., Gredebäck, G., Koch, B., & Rohlfing, K. J. (2016). Language meddles with infants' processing of observed actions. *Frontiers in Robotics and AI*, 3, Article 46.
- Singh, A., & Rohlfing, K. J. (2024). Coupling of task and partner model: Investigating the intra-individual variability in gaze during human–robot explanatory dialogue. In *Proceedings of the International Conference on Multimodal Interaction (ICMI Companion '24)*. San Jose, Costa Rica: ACM.
- Skordos, D., Bunger, A., Richards, C., Selimis, S., Trueswell, J., & Papafragou, A. (2020). Motion verbs and memory for motion events. *Cognitive Neuropsychology*, 37(5–6), 254–270.
- Slobin, D. I. (1996). From “thought and language” to “thinking for speaking.” In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). New York: Cambridge University Press.
- Stern, G. (1931). *Meaning and change of meaning; With special reference to the English language*. Wettergren & Kerbers.
- Tagliani, M., Vender, M., & Melloni, C. (2022). The acquisition of negation in Italian. *Languages*, 7(2), 116.
- Talmy, L. (1975). Semantics and syntax of motion. In J. P. Kimball (Ed.), *Syntax and semantics Volume 4* (pp. 181–238). Brill.
- Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description. Volume III: Grammatical categories and the lexicon* (pp. 57–149). Cambridge: Cambridge University Press
- Talmy, L. (2000). *Toward a cognitive semantics*, volume 2 of *Series 44*. Cambridge, MA: MIT Press.
- Tettamanti, M., Manenti, R., Della Rosa, P. A., Falini, A., Perani, D., Cappa, S. F., & Moro, A. (2008). Negation in the brain: Modulating action representations. *NeuroImage*, 43(2), 358–367.
- Tian, Y., & Breheny, R. (2016). Dynamic pragmatic view of negation processing. In P. Larrivée & C. Lee (Eds.), *Negation and polarity: Experimental perspectives* (pp. 21–43). Cham: Springer International Publishing
- Tian, Y., Ferguson, H., & Breheny, R. (2016). Processing negation without context - Why and when we represent the positive argument. *Language, Cognition and Neuroscience*, 31(5), 683–698.
- Tomasino, B., Weiss, P. H., & Fink, G. R. (2010). To move or not to move: Imperatives modulate action-related verb processing in the motor system. *Neuroscience*, 169(1), 246–258.
- Wason, P. C. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior*, 4(1), 7–11.
- Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between l1 and l2 speakers of English. *Journal of Phonetics*, 70, 86–116.
- Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. *WIREs Cognitive Science*, 2(3), 253–265.
- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1), 3–36.
- Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1), 1–34.

- Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science*, 11(1), 73–77.
- Wrede, B., Schillingmann, L., & Rohlfing, K. (2013). Making use of multi-modal synchrony: A model of acoustic packaging to tie words to actions. In M. Spranger & K.-U. Kühnberger (Eds.), *Theoretical and computational models of word learning: Trends in psychology and artificial intelligence*. (pp. 224–240). Atlantis Thinking Machines Paris: Atlantis Press.
- Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., & Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, 4(6), 651–655.
- Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, 16(2), 80–84.
- Zacks, J. M., Swallow, K. M., Vettel, J. M., & McAvoy, M. P. (2006). Visual motion and the neural correlates of event perception. *Brain Research*, 1076(1), 150–162.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1: Estimates for post-hoc pairwise comparison for recall (Experiment 1)

Table S2: Estimates for post-hoc pairwise comparison for recall (Experiment 2)

Table S3: Experiment 1—Classification of errors made during recall when an incorrect subaction was identified

Table S5: Random effects estimates and plots showing recall by subject and by item for Experiment 1

Table S4: Experiment 2—Classification of errors made during recall when an incorrect subaction was identified

Table S6: Random effects estimates and plots showing recall by subject and by item for Experiment 2

Table S7: Random effects estimates and plots for goal-looks by subject and by item for action types in Fig. 2: C1 (Action A), C2 (Action B), C3 (Action C), and C4 (Action D).

Figure S3: GAM difference smooth comparison between no-instruction and other instruction conditions as reported for Figs. 3 and 5 in the paper

Manners Matter: Action history guides attention and repair choices during interaction

Amit Singh^{1*}, Britta Wrede^{2,3}, Birte Richter^{2,3}, André Groß^{2,3}, Katharina J. Rohlfing¹

Abstract—This study investigated how action histories – unfolding sequences of actions with objects – provide a context for both attentional allocation and linguistic repair strategies. Building on theories of enactive cognition [34] and sensorimotor contingency theory [40], we experimentally manipulated action sequences (action history) to create either simple or rich “situational models,” and investigated how these models interact with attention and reflect in linguistic processes during human-robot interaction. Participants ($N = 30$) engaged in a controlled object placement task with a humanoid robot, where the action (manner) information was either provided or omitted. The omission elicited repair behaviors in participants that were in focus of our investigation. For rich models (competing action possibilities) participants demonstrated: a) increased attentional reorientation, reflecting active engagement with the situational model b) preference for restricted repairs, targeting the specific source of trouble in action selection. Conversely, a simple situational model led to more generalized attention patterns and open repair strategies, suggesting weaker constraints on internal processing. These findings highlight how situational structures emerge externally to scaffold internal cognitive processes, with action histories serving as a crucial context for the interface between perception, action, and language. We discuss how to implement such a tight loop in the assistance of a system.

Index Terms—Attention, Action, Repairs, Task model, HRI, Eyemovements

I. INTRODUCTION

At any given moment in interaction, human behavior is dynamically shaped by interplay between external information presented in the environment, let’s say: a situation model [28], [39], [41], [47], [49], and internal simulations of historical context, generated through experience, let’s say: a task model [1], [12], [29]. Externally, a situation model represents the immediate perceptual and contextual information available [27], [39], [41]. Internally, behavior is influenced by a task model, a simulated history of actions and goals, often mediated through language [17], [35], [58]. These external and internal states are brought together—a moment that was described as “situatedness” [42, pp. 132]. The function is to simultaneously interpret the aspects of the world and prescribe possible actions that can be chosen to achieve the goal [11]. In a more radical stance, this *enactive*

notion has been dubbed as “seeing as a way of acting”, [40, pp. 939], where cognitive processes do not merely serve to passively create a “world model” around us, but instead proactively ascribe what actions are possible within a given situation [18]. Thus, sensory processes are always, and to a considerable degree, influenced by conceptualized possible actions [11], [18], [52]. Consistent with this idea, several studies have shown that action preparation leads to a shift of attention toward object-related information [9], be it object’s feature or it’s function. Notably, object function has been particularly shown to offer a faster recognition than object features [9], suggesting that action-relatedness may serve as a key constraint guiding attention in real-world interactions. This notion is further supported by sensorimotor contingency theory [40], for example, when we look at an object, our brain processes visual information not just based on what we see (features), but also based on how we act upon it (possible actions), where the latter is often precursor of sensory processes guiding performance [44]. Following upon this, extensive research has explored both the perspectives: feature [51] and action based attention [20] also known as “action plans” [19, pp. 144]. Whereas considerable attention has been given to these two perspectives, limited focus has been placed on understanding how action history might act as a context for attention and language, particularly in situations where past actions are characterized by *manners* that influence how a subsequent action should be performed. This gap is particularly notable, as an object-bound-manner conceptualizes *how* we should interact with an object, shaping the internal task model [9], [20], in contrast to features that pertain to external constraints and object semantics [51]. Thus, an object-bound-manner can also have propensity for guiding attention, albeit at sensory level or in often cases might precede it. When it comes to foraging for information, on one hand, the richness of information that surrounds us at any given moment presents overwhelming and boundless possibilities for action; on the other hand, our cognition seem to have evolved to operate parsimoniously – by optimizing attention [27], [47] and employing language in the most efficient way possible [33]. When faced with uncertainty, our cognition dynamically shifts both outward and inward, relying on external situational cues available in the moment and drawing on internal experiences [31], [56]. Take, for example, a conversation where misunderstanding arises. Our cognition dynamically works to resolve it by exploiting the *immediate context* and utilizing the *past experiences*, collaboratively engaging with the conversational partner to reach a solution. Crucially, in this case, the partner is communicated

*This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/I 2021-438445824 “Constructing Explainability”.

¹Psycholinguistics, Faculty of Arts and Humanities, Paderborn University, Germany.

²Medical Assistance Systems, Medical School OWL and ³Center for Cognitive Interaction Technology, CITEC from Bielefeld University, Germany. *Corresponding Author: Amit Singh, amit.singh@uni-paderborn.de

of problem source, through linguistic (e.g., repair initiation) or non-linguistic means [13].

These processes serving each other manifest at both attention – where individuals selectively reorient for information [22] – and linguistic level, where choices such as a particular repair strategies are preferred over other in case of misunderstanding [15]. Whereas theories suggest how the external environment can function as an *external memory* in interaction [56, pp. 628], there is little explanation for how action history guides internal attention, particularly when it encapsulates object-bound manners. A promising framework for turning ‘inside’, whereby modeling attention in such a scenario is proposed by Theory of Visual Attention (TVA) [3], which not only accounts for historical influences but also considers the processing resources available at a given moment in interaction, known as capacity. It has been shown that specific linguistic constructions, such as negations decrease capacity in dual-task paradigm [7]. However, it was shown that negation changed the distribution of attention as opposed to an affirmation in such a way, that both objects in a scene received the same amount of visual attention. This idea has been implemented in a computational model to account for available cognitive resources online in interaction and provide a pathway to account for the reduction of capacity [23]. More specifically, this approach uses a generalized concept of capacity that is based on the observed decrease of visual capacity after encountering a negation. Using a scoring scheme based on a range of monitoring factors (interaction history, task understanding) linguistic scaffolding means, such as negations, are determined linearly. In this case, if a negation has been used by the robot in the prior interaction history, it is assumed that the capacity of the listener is reduced, hence appropriate scaffolding means are selected. More recently, this model has been evaluated in a user study showing that, although processing cost following a negation is increased (as measured by reaction time), task performance can significantly be improved when such linguistic scaffolding strategies are applied appropriately [24]. Whereas in the above-mentioned model, cognitive capacity is influenced by instruction (negation) participants receive from the system, information held in the memory during a task such as action history might also influence the capacity and overall resources left during interaction. Specifically, the way objects can be acted upon – including their possible affordances – can shape both cognitive capacity and resources available during an interaction, which may, in turn, be reflected in linguistic behavior. To bring the aforementioned ideas in a unified framework, we needed, a) a formal approach to ground actions in these theories, b) a real-world interaction study to further investigate it. An advantage of such *situated* and context aware models are their utility for learning and understanding, which addresses the fundamental question of how attention prioritizes information by negotiating between history and information readily offered by the environment. Such models might also play a crucial role in developing user-centered individual guidance which can assist users and promote their learning or action performance. Hereby, we

take advantage of the user’s behavior: Their repair encoding strategies, serving as a gateway to the task model—a snapshot of an explainee’s current understanding of the task, and what possible action (s)he is currently focusing and thinking about. This assumption not only incorporates simulated history but also appreciates the immediate contextual information available externally. Given the complexity of this issue, we start by outlining the key elements of a *situation* emerging from actions.

A. Situational models

Whereas from the computational science, situation consists of facts that are physically (or socially) given [43], from the human science, a situation is always a construction binding what is present with experience to make sense of it [4]. This is also the reason for why the terms *situation* and *context* “are often used synonymously” [42, 133]. A situation can be defined as “spatiotemporal ordering of objects and agents alongside physically given constraints or characteristics like gravitational force or light intensity” [42, 133] and it has the potential to activate specific behaviors [5]. For example, the availability of a hook makes actions of hanging on it possible. While this example seems simple, identifying all parameters that sum up to a situation in an objective manner is not trivial. For humans, relevant parameters are created in the moment of acting within the situation by considering external circumstances and internal interpretation of them, retrieved with the aid of memory. A situation also provides opportunity to an agent within it to refer back to the physical (or social) facts readily available, often and even though when it is held internally in the memory [16], [27]. For instance, it has been found that participants, although hold a “task model” in working memory after turning away from the environment where the object is still placed, refer and reorient themselves back even in low complexity tasks such as object-copying [2], [16]. In an obvious case, participants reorientation occurs frequently when some misunderstanding occurs, or insufficient information provided [45], which is also shown to be a function of task complexity and relative distance between source and goal locations [2], [16]. It seems that an internal task model operates on the basis of a context consisting of already performed actions. Recent study shows that when participants perform sequential actions instructed by a robot, in cases where insufficient information is provided, participants first attend to the task and try to solve it based on the current task model and history. This suggests that the primary locus of attention is always the information readily available in front [45]. If this is not sufficient, language can solicit the missed information. In this study, a simplistic action model was used and consisted of the elements *path* and *manner*. Whereas *path* refers to the location of an object *where* it should be placed, the *manner* refers to the way *how* the objects should be placed.

B. Repairs: a window to the task model?

As suggested in the previous section, communication serves as a means of disambiguation, requiring participants to

retain some task model. Although the task model is directly inaccessible, one way to investigate it is by analyzing the linguistic processes that emerge precisely when the task model is needed. Consider a scenario, where a child is trying to solve a puzzle and a concerned piece has to be fit by rotating it. The child will initially rely on her current task model, shaped by past experiences and action for a given situation. If she struggles to decide for the current action independently, a question word might follow to the caregiver. She might ask an open-ended question: ‘How?’. However, given the situational constraints, if the child has consistently used a specific manner up to that point, her task model might be more restricted toward assuming the re-use of this manner, leading to a linguistic construction like ‘should I join it that way?’. This shift reflects how linguistic repairs target the source of misunderstanding in different ways—either through open-ended requests or more restricted request, precisely targeting the specific manner in action repertoire [15]. Thus, repairs may provide a window to the task model emerging from the history in an interaction, thereby serving a functional purpose [15]. Motivated by a longstanding literature, here, we use repairs at two dimensions to formalize our task model: open and restricted. Whereas open repair types are said to leave open the source of problem and requires a *general* hold on the history, restricted type point to specific aspect of the trouble source of the previous actions, and thus subsumes the fine-grained availability of the possible actions in the action repertoire [14, 107].

II. CURRENT STUDY

To ground actions in the provided theories, we followed our previous setup involving real-world interactions between humans and a robot [45]. The interactive nature of this setup not only allowed for the emergence of repair processes but also provided a systematic way to manipulate the situation model by varying action components: *path* and *manner*. Human participants, acting as explainees, could ask the robot questions. Previous research has shown that this setup is effective in scenarios where participant-driven manner manipulation is required [26], [53]. Unlike display-based experiments, this setup allows the use of real objects as explananda, enabling participants to physically interact by following the robot’s guidance. Since the task simply involves sequentially placing objects in one of two possible manners, the accumulation of these actions over time forms a situation model shaped by action history. A key distinction between the two example situation models lies in their action history: In one case, only *one* action manner is present in the working memory before the final action takes place (Fig. 1, top), while in the other, *two* manners are already part of the action repertoire (Fig. 1, bottom). The final action, therefore, builds upon an experience with either only a single manner (simple situational model) or a combination of two (rich situational model), influencing what task model emerges.

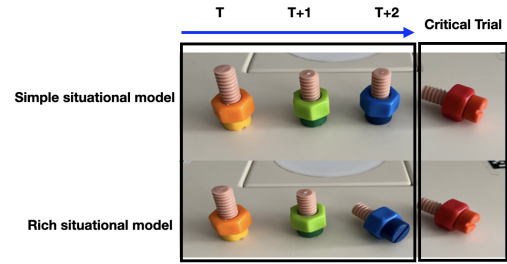


Fig. 1. An example of two different situational models made up of a sequence of actions: simple (Top) and rich (Bottom). Time points labeled T, (T+1), and (T+2) correspond to the action history, before the critical trial is presented.

A. Predictions

Our prediction relates to first assessing attention operating for two situational models (rich and simple, as described above), and then examining how the internal task model is reflected in the linguistic processes, such as repairs. Crucially, we were interested in the time before the repair utterance, since this is the time where linguistic encoding takes place [32], [36]. When under-information is encountered in the task, responses may emerge at both the attentional and linguistic levels. These responses could involve both an attempt to resolve the task by relying on the established task model, shaped by the emerging context (situational models), and turning to the explainer to reveal the task model through linguistic means, such as repair requests (as discussed in section I-B). These two processes – context-driven problem-solving and linguistically-driven clarification – should be reflected in both attention patterns and linguistic behavior. Based on these assumptions, we formulated specific expectations for how attention and linguistic responses might vary across situational models.

Situational models drawing attention: When there is under-information associated with the final action (i.e., the robot omits the information about action manner), an obvious strategy is to decide for the possible action by exploiting the given situation and simulated history. According to the sensorimotor contingency theory, such simulations should also be reflected in the eye movements and at the sensory level [40]. A rich situational model not only provides an opportunity to externally refer to the physical facts when encountered with underinformativeness, but also serves for an internal task model where all the possible manners are available [47]. Thus, we expected a rich situational model to engage more attention than simple situational model, and hence promoting more external reliance.

Repairs reflecting task model: Importantly, the internal simulation of actions toward a goal (task model) should be reflected at the linguistic level. It is suggested that a restricted request presumes more fine-grained understanding of the possible actions due to its requirement for precisely targeting

the source of trouble [13]. To do so, all aspects of the action must be present in the task model—something more likely to happen in rich situational models where all possible actions are available. Consequently, we expect restricted repairs of the kind “Should I place it vertically or horizontally?” to occur more frequently in rich situational models than in simple situational models. Based on a parsimonious account, which suggests that open requests are less cognitively demanding [14], we expected to observe open requests (e.g., “How?”) more frequently than restricted requests overall.

III. METHOD

We employed same methodology as in our previous experiment investigating gaze behavior in human–robot explanatory dialogue [45]. The methods are approved by Review Board of the university, and informed consent was obtained from all participants prior to their involvement. The study was preregistered on Open Science Framework (OSF) before data collection commenced ¹, where the analysis codes are publicly available, and the data will be made accessible upon request.

A. Setup

As depicted in Figure 2, participants were seated facing the NAO [21] robot while wearing Pupil Core [30] eye-tracking glasses from Pupil Labs, which operated at a sampling rate of 120 Hz. A 5-point calibration procedure was done, achieving an accuracy of 0.60° and a precision of 0.02°. The experimental environment was carefully controlled for consistent illumination and noise isolation. A transparent glass container, placed on a table in front of the participant, held the objects used in the task. Alphabet stickers were arranged in a row to serve as the designated destination path for the objects. Two stationary cameras, positioned on either side of the setup, recorded the experimental proceedings. To minimize potential biases in interaction between robot and participant, an experimenter zone was established behind the participant, concealed by a divider. This arrangement ensured that experimenter remained out of view, allowing for an uninterrupted interaction exclusively between robot and participant. The experiment commenced with calibration of the eye-tracking system, followed by an introduction to the task.

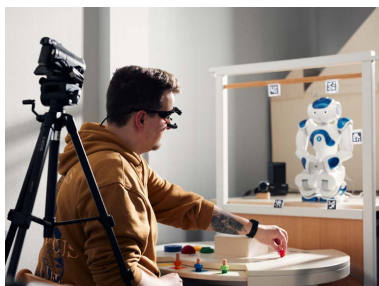


Fig. 2. Figure depicting the experimental setup

¹Project link: osf.io/6xc7u

B. Participants

Thirty-three university students (mean age = 22.30, $SD = 2.63$) were recruited through classroom advertisements and flyer distribution. All participants demonstrated native or fluent proficiency in German, ensuring comprehension of instructions during interaction. Data from three participants were excluded from analysis due to track-loss issues in the eye-tracking sample points, resulting in a final sample size of thirty participants.

C. Stimuli

The stimuli consisted of four screw-shaped objects, each in a distinct color: green, red, yellow, and blue (Fig. 3). Each object could be placed in a specific *manner*: vertically or horizontally (termed: “hochkant” or “quer” in German, respectively) and on a destined but random location (indicated by letters) at time T, T+1, T+2 together forming a context before the critical trial is presented (Fig. 1). To maintain task complexity, manner options were limited to only two affordances (vertical or horizontal). The combination of manners and different locations (indicated by letters) was randomly generated for each trial. This provided us with two conditions: In condition with a simple situational model, all manners were the same (either all vertical or all horizontal; in condition with a rich situational model, two manners were represented (Fig. 1). Participants were engaged with these two conditions, before the final full-information or critical trail – with omitted manner – was randomly introduced. The randomization ensured not only occasional omission of manners prompting participants to initiate repairs, but also provided diverse array forming two situational models.

D. Procedure

The robot was programmed with semi-autonomous capabilities, allowing the experimenter to take control in cases where the robot failed to respond or the conversation broke down. The robot began by introducing itself and reiterating the task instructions. After confirming the participant’s readiness, the robot commenced the first trial. Following each instruction, the robot briefly gazed at the task area and then at the participant, seeking confirmation: “Hast du das erledigt?” (“Have you done it?” in English). Upon receiving affirmation, the robot proceeded to the next object. A trial concluded once all four objects were placed according to the specified manner on a location. The robot then instructed participants to return the objects to their original positions and turn away to solve a simple arithmetic task (e.g., addition or subtraction). This step was designed to offload working memory and maintain participant engagement, preventing passive task-following. After completing the arithmetic task, participants turned back to the robot and performed a recall task, replicating the object placements as instructed during the trial. Once confirmed, the robot advanced to the next trial. The experiment consisted of four blocks. Each block included five trials, with each trial involving four object placements, resulting in a total of 20 moves per block. Moves were randomized

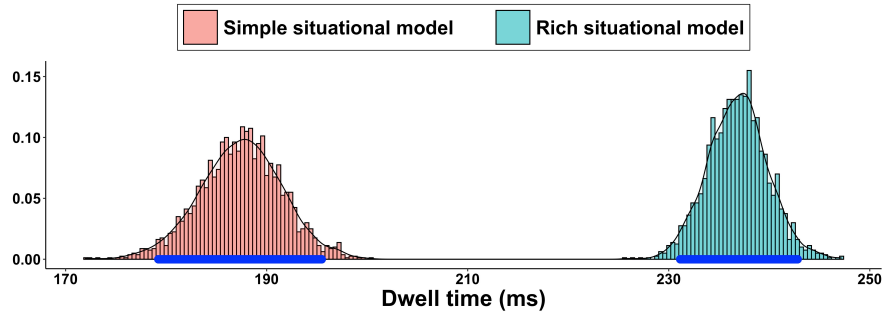


Fig. 3. Posterior predictive estimates of dwell times for the two situational models. Blue horizontal lines depict the 95% CI.

within each block where information regarding the path or manner was intentionally omitted in four randomly selected moves (missing information trials), while the remaining 16 moves provided complete information (full information trials). Higher full-information trial and occasional missing information trial was introduced to minimize exhaustion and prompt participants to seek clarification only when needed. The entire experimental session lasted approximately 45 minutes, during which eye-gaze data and video recordings were continuously collected.

Instruction to the participants: Before the experiment started, each participant was instructed to follow the robot's guidance for placing each object and was encouraged to ask questions if they felt that the provided information was insufficient. Participants were instructed to repeat their commands, in case they feel the robot was unable to process their input. Since the missing information pertained only to the path or the manner of placement, we anticipated that participants would frame their questions along these dimensions (e.g., "Where should I place it?" or "How should I position it?"). However, participants were also given the freedom to request repetitions of instructions whenever they deemed it necessary. Each trial involved a specific instruction for placing an object in a designated manner and along a specific path. An example instruction template is provided as follows:

Place the green object [Manner] on [Path].

Lege bitte das grüne Objekt [Hochkant or Quer] auf [A].

E. Analysis

Our analysis was predominantly focused on the processes involved during the critical trial, where the manner information was missing. We specifically focused on the manner than the path component, since it has been shown previously that manner learning might require more scaffolding than the path [26]. Figure 4 provides an overview on the repairs and their type. Rectangular areas of interest were drawn around each location surrounding the object in which fixations were computed within a time window of 80 to 350 ms, using a maximum dispersion set default at a value of 1.5 degrees. To classify the request types in open and restricted types, we followed the method described in [14]. The request

TABLE I
POSTERIOR ESTIMATES FOR SITUATIONAL MODELS

Term	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	212.07	1.74	208.69	215.47
Context	49.47	3.51	42.55	56.31

TABLE II
PARAMETER ESTIMATES FOR REQUEST TYPES

Term	Estimate	Std. Error	z-value	p-value
Intercept	-1.92	0.74	-2.60	0.009
Context:Rich	1.94	0.50	3.85	<0.001

types where either of the manners were mentioned were classified as restricted, and general requests without manner were classified as open type. We considered the time from end of critical trial till the onset of the repair, providing us the precise window of linguistic encoding. These time-windows along with the repair types were annotated by 2 student assistants in ELAN software [55], who were unaware about the predictions of the study (refer, II-A). For the first set of analysis investigating the effect of situational models on attention, we used a Bayesian approach. In this analysis, we estimated the effects of situational models on attention by incorporating prior distributions and updating them with the observed data to derive posterior estimates. For the subsequent analysis examining the effects of question types, we utilized a Generalized Linear Mixed Effects Model (GLMER) [8], which allowed us to account for both fixed and random effects in the data. All analyses were conducted in R [38].

IV. RESULTS

Situational models drawing attention: We tested our first prediction regarding two situational models affecting attention. A Bayesian model was fitted on gaussian likelihood and mildly informative priors for intercept and slope². The model was coded in brms: Bayesian Regression modeling using Stan [6] a front-end for the probabilistic programming language Stan [48]. We modeled dwell time

²Prior specifications: $\alpha \sim \text{Normal}(200, 50)$, $\beta \sim \text{Normal}(0, 100)$, $\sigma \sim \text{Normal}^+(50, 10)$.

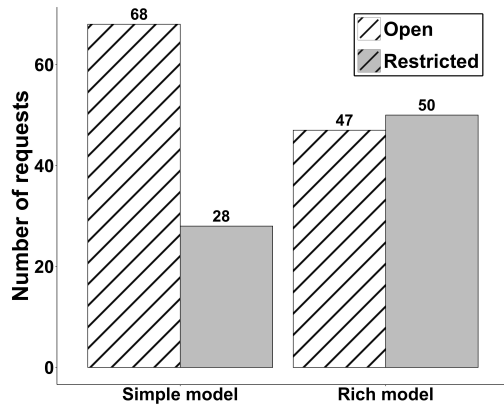


Fig. 4. Repair types across two situational models

as a function of the situational models with two levels (simple vs. rich). The independent variable (situational model) was contrast-coded using a sum-to-zero coding scheme, where +0.5 was assigned to rich and -0.5 to simple, ensuring that the levels are centered around zero. The posterior estimates are shown in Table I and the density is visualized in the Figure 3. It was found that a rich situational model was drawing more attention than simple situational model when participants were faced with underinformativeness. This was particularly shown by the slope coefficient for dwell times across both models ($\beta = 49.47$, $SD=3.51$, $CI[42.55, 56.31]$). A bayes factor analysis further confirmed a very strong evidence in favor of the effect of situations ($BF = 7414$), which supported our first prediction regarding situational models modulating attention.

Repairs reflecting task model: Here, we tested our second prediction whether the task model emerged from rich and simple situations also guided the selection of repairs. We modeled the repair types as a function of situation (rich vs. simple). Since the request types were categorical variables, we transformed it to numerical (restricted = 1, open = 0) and modeled it with binomial assumption. Here, the intercept parameter revealed the log-odds of the restricted request and slope parameter the difference between situations (rich vs. simple). We fitted a glmer model by accounting for individual subject variability by including by-subject adjustment to the intercept as random factor. The model parameters are provided in the Table II. We first analyzed the overall distribution of request types for both the situational models (see Fig. 4). We found a significant main effect of repair type, such that overall restricted requests were lower than open ones ($\beta = -1.92$, $SE=0.74$, $p<.01$). This suggests that participants were more likely to use open requests that did not target any specific manner, regardless of situation models. Furthermore, there was a significant interaction between repair type and situational model ($\beta=1.94$, $SE=0.50$, $p<.001$), suggesting that participants were more likely to use restricted request in presence of rich situational model than simple situational model.

V. DISCUSSION

In present study, we investigated how sequences of actions manifest situational models that guide attention and influence linguistic processes during human–robot interaction. The emergence of situational models – rich, which encapsulate two possible action manners, and simple pertaining to only one action manner before the final action is performed – provides a unique history and context for the task model. Whereas we conceptualized the situational model as an entity readily formed and available ‘externally,’ the task model is a product of mental processes, revealed through linguistic means. We motivated our investigation from theory of enactive cognition [34] and sensorimotor contingency [40] and found that attention is driven by an internal action simulation contingent upon the external situational model, particularly when this model offers a rich affordance. This was evident by a higher dwell time for rich situational model than simple situational model (Fig. 1). Although the idea that situational models function as a deictic mechanism for memory is not new and has been demonstrated in multiple studies [31], [56], particularly in tasks relying on working memory [16] or in interactions without a specific task [2], we propose that these models may also emerge from sequences of actions building upon each other, allowing them to be referenced when uncertainty arises in the task. Previous studies have suggested that this reference serves an ecological purpose by offloading and delegating internal task demands to external sources, given the limited capacity of our working memory [1]. Our study highlights that this delegation might also depends upon nature and richness of the situational constraints serving our enactive abilities. Unlike previous studies, which focused on ‘what’ and ‘where’ features to characterize the external model, appreciating the enactive notion, we demonstrate that actions (and their manners) related to *how* might also be an integral part of its nature. The findings also reveal that the type of linguistic repair chosen by participants is influenced by the situational model (Fig. 4). This influence might reflect the participant’s task conceptualization (task model), as language production – unlike comprehension – relies on a controlled process that requires conscious cognitive operations [32], [36]. More specifically, we found that a rich situation enabled the participants to use restricted repairs to highlight the precise trouble source, which requires a fine-grained understanding of possible actions. Research on turn-taking supports that restricted requests may require more cognitive resources than open request due to their ability to target the ‘local’ trouble source of the previous turn—a behavior that we observed and justified by the parsimonious account (i.e., less cognitive load) with open requests overall higher than restricted requests (Fig. 4). Remarkably, restricted requests occurred more frequently for rich situational model. This suggests that although restricted requests are cognitively demanding, a rich situational model may provide a context for their efficient use, particularly when all possible manners were available in the action repertoire. Following this, a rich situational model may free

up cognitive resources, making it more plausible to process cognitively demanding (but helpful) linguistic constructions for scaffolding, such as negations [10]. This plausibility has already been demonstrated in a previous study in pragmatics [54]. Wason [54] suggests that processing cost associated with negation can be mitigated, when all competing features are presented in the context. Here, we proposed it as a rich situational model and extend its utility for actions. Hence, a rich situational model might provide an advantageous context for implementing and testing assumptions for the interactive models employing negation, where the capacity is posing a challenge. For example, in a model such as the SHIFT [23], the participant's current and next capacity in two different situations can be estimated based on the linguistic behavior to select the appropriate linguistic strategy, such as negations or assertions. Similarly, in a reinforcement learning approach as proposed in SHIFT model [23], linguistic behavior could be used as a reinforcement signal indicating whether a prior strategy has increased or decreased the attentional capacity. More specifically, a participant asking an open question in a rich situation indicates low capacity whereas a restricted question in a rich situation would indicate a high capacity. For a computational model, these results provide deep insights that can be drawn from cues that are easily derivable from the verbal behavior of the participant. Whereas existing approaches such as proposed in [23] only provide a prior estimate of the capacity based on the use of scaffolding means such as negation that are difficult to process, the analysis of repairs in specific situations can shed light on the granularity of the internal simulation and its components. For example, an open repair in a rich situation may indicate, (a) low capacity and/or (b) only partial processing of the rich situation containing two manners. This, in turn, provides insights as to what scaffolding strategies to apply to better assist the human in performing actions. More research is required to determine which scaffolding strategies are beneficial in this situation. While our study provides insights into interplay between situational models and emerging task models, on which attention and language behaviors are based, there are limitations that we aim to address in future research. First, our task was relatively simple, involving only two manners. Future studies could explore more complex tasks with multiple manners to investigate its generalization, particularly given that manners are fundamental aspect of actions as acknowledged in psycholinguistic research [46], [50]. Second, although our predictions were motivated by studies on human–human interaction, the validity is yet to be tested with human interlocutors. In conclusion, our study demonstrates that action history (experience with action sequences) provides a situational and task model, guiding attention and linguistic repair strategies during human–robot interaction. We integrated insights from cognitive science, psycholinguistics, and human–robot interaction to show how external situational and internal task models may interact to shape human behavior and inform the design of adaptive systems that can then assist humans in performing actions.

ACKNOWLEDGMENT

We thank our participants and student assistants Annabell Horch, Lasse Moritz Möller, and Tim Schaffhauser.

REFERENCES

- [1] A. Baddeley, "The episodic buffer: a new component of working memory?," *Trends Cogn. Sci.*, vol. 4, no. 11, pp. 417–423, 2000.
- [2] D. H. Ballard, M. M. Hayhoe, and J. B. Pelz, "Memory representations in natural tasks," *J. Cogn. Neurosci.*, vol. 7, no. 1, pp. 66–80, 1995.
- [3] C. Bundesen, "A theory of visual attention," *Psychol. Rev.*, vol. 97, no. 4, p. 523, 1990.
- [4] J. R. Binder and R. H. Desai, "The neurobiology of semantic memory," *Trends Cogn. Sci.*, vol. 15, no. 11, pp. 527–536, 2011.
- [5] P. Brézillon, "Context in problem solving: A survey," *Knowl. Eng. Rev.*, vol. 14, no. 1, pp. 47–80, 1999.
- [6] P.-C. Bürkner, "brms: An R package for Bayesian multilevel models using Stan," *J. Stat. Softw.*, vol. 80, no. 1, pp. 1–28, 2017.
- [7] N. C. Banh, J. Tünnermann, K. J. Rohlfing, and I. Scharlau, "Benefiting from binary negations? Verbal negations decrease visual attention and balance its distribution," *Front. Psychol.*, vol. 15, p. 1451309, 2024.
- [8] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.
- [9] M. S. Castelano and R. L. Witherspoon, "How you use it matters: Object function guides attention during visual search in scenes," *Psychol. Sci.*, vol. 27, no. 5, pp. 606–621, 2016.
- [10] H. H. Clark and W. G. Chase, "On the process of comparing sentences against pictures," *Cognit. Psychol.*, vol. 3, no. 3, pp. 472–517, 1972.
- [11] A. Clark, *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press, 1998.
- [12] L. Davachi and S. DuBrow, "How the hippocampus preserves order: the role of prediction and context," *Trends Cogn. Sci.*, vol. 19, no. 2, pp. 92–99, 2015.
- [13] M. Dingemanse, S. G. Roberts, J. Baranova, J. Blythe, P. Drew, S. Floyd, et al., "Universal principles in the repair of communication problems," *PLoS ONE*, vol. 10, no. 9, p. e0136100, 2015.
- [14] M. Dingemanse and N. J. Enfield, "Other-initiated repair across languages: towards a typology of conversational structures," *Open Linguist.*, vol. 1, no. 1, 2015.
- [15] M. Dingemanse and N. J. Enfield, "Interactive repair and the foundations of language," *Trends Cogn. Sci.*, vol. 28, no. 1, pp. 30–42, 2024.
- [16] D. Draschkow, M. Kallmayer, and A. C. Nobre, "When natural behavior engages working memory," *Curr. Biol.*, vol. 31, no. 4, pp. 869–874.e5, 2021.
- [17] D. Draschkow and M. L.-H. Vo, "Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search," *Sci. Rep.*, vol. 7, no. 1, p. 16471, 2017.
- [18] A. K. Engel, A. Maye, M. Kurthen, and P. König, "Where's the action? The pragmatic turn in cognitive science," *Trends Cogn. Sci.*, vol. 17, no. 5, pp. 202–209, 2013.
- [19] F. van Ede and A. C. Nobre, "Turning attention inside out: How working memory serves behavior," *Annu. Rev. Psychol.*, vol. 74, pp. 137–165, 2023.
- [20] J. J. Gibson, *The Ecological Approach to Visual Perception: Classic Edition*. New York, NY: Psychology Press, 2014.
- [21] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "Mechatronic design of NAO humanoid," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 769–774.
- [22] I. C. Griffin and A. C. Nobre, "Orienting attention to locations in internal representations," *J. Cogn. Neurosci.*, vol. 15, no. 8, pp. 1176–1194, 2003.
- [23] A. Groß, B. Richter, and B. Wrede, "SHIFT: An interdisciplinary framework for scaffolding human attention and understanding in explanatory tasks," *arXiv*, 2025, Art. no. 2503.16447.
- [24] A. Groß, B. Richter, B. Thomzik, and B. Wrede, "Leveraging cognitive states for adaptive scaffolding of understanding in explanatory tasks in HRI," *arXiv*, 2025, Art. no. 2503.19692.
- [25] A. Groß, A. Singh, N. C. Banh, B. Richter, I. Scharlau, K. J. Rohlfing, and B. Wrede, "Scaffolding the human partner by contrastive guidance in an explanatory human–robot dialogue," *Front. Robot. AI*, vol. 10, p. 1236184, 2023.

- [26] A. Groß, A. Singh, N. C. Banh, B. Richter, I. Scharlau, K. J. Rohlfing, and B. Wrede, "Scaffolding the human partner by contrastive guidance in an explanatory human-robot dialogue," *Front. Robot. AI*, vol. 10, 2023.
- [27] M. M. Hayhoe, D. G. Bensinger, and D. H. Ballard, "Task constraints in visual working memory," *Vision Res.*, vol. 38, no. 1, pp. 125–137, 1998.
- [28] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends Cogn. Sci.*, vol. 9, no. 4, pp. 188–194, 2005.
- [29] P. N. Johnson-Laird, "Mental models," in *Foundations of Cognitive Science*, M. I. Posner, Ed. Cambridge, MA: MIT Press, 1989, pp. 469–499.
- [30] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 1151–1160.
- [31] D. Kirsh and P. Maglio, "On distinguishing epistemic from pragmatic action," *Cognit. Sci.*, vol. 18, pp. 513–549, 1994.
- [32] W. J. M. Levelt, *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press, 1989.
- [33] S. C. Levinson, *Pragmatics*. Cambridge, UK: Cambridge Univ. Press, 1983.
- [34] M. McGann, "Enactive theorists do it on purpose: Toward an enactive account of goals and goal-directedness," *Phenomenol. Cogn. Sci.*, vol. 6, pp. 463–483, 2007.
- [35] A. Papafragou, J. Hulbert, and J. Trueswell, "Does language guide event perception? Evidence from eye movements," *Cognition*, vol. 108, no. 1, pp. 155–184, 2008.
- [36] M. J. Pickering and S. Garrod, "An integrated theory of language production and comprehension," *Behav. Brain Sci.*, vol. 36, no. 4, pp. 329–347, 2013.
- [37] R. Pulverman, R. M. Golinkoff, K. Hirsh-Pasek, and J. S. Buresh, "Infants discriminate manners and paths in non-linguistic dynamic events," *Cognition*, vol. 108, no. 3, pp. 825–830, 2008.
- [38] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Found. Stat. Comput., 2023.
- [39] J. K. O'Regan, "Solving the 'real' mysteries of visual perception: the world as an outside memory," *Can. J. Psychol./Rev. can. psychol.*, vol. 46, no. 3, p. 461, 1992.
- [40] J. K. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness," *Behav. Brain Sci.*, vol. 24, no. 5, pp. 939–973, 2001.
- [41] R. Ramos Gameiro, K. Kaspar, S. U. König, et al., "Exploration and exploitation in natural viewing behavior," *Sci. Rep.*, vol. 7, p. 2311, 2017.
- [42] K. J. Rohlfing, M. Rehm, and K. U. Goecke, "Situatedness: The interplay between context(s) and situation," *J. Cogn. Cult.*, vol. 3, no. 2, pp. 132–156, 2003.
- [43] K. J. Rohlfing, K. Främling, and F. Kern, "Context for explanations," in *Handbook of Social Explainable AI*, K. J. Rohlfing, K. Främling, S. Alparsancar, K. Thommes, and B. Y. Lim, Eds. (in press).
- [44] W. X. Schneider and H. Deubel, "Selection-for-perception and selection-for-spatial-motor-action are coupled by visual attention: A review of recent findings and new evidence from stimulus-driven saccade control," in *Common Mechanisms in Perception and Action*, vol. 19, 2002, pp. 609–627.
- [45] A. Singh and K. J. Rohlfing, "Coupling of task and partner model: Investigating the intra-individual variability in gaze during human-robot explanatory dialogue," in *Proc. 26th Int. Conf. Multimodal Interact. Companion (ICMI Companion '24)*. New York, NY, USA: ACM, 2024, pp. 218–224.
- [46] D. Slobin, "Thinking for speaking," in *Proc. 13th Annu. Meet. Berkeley Linguist. Soc.*, 1987, pp. 435–444.
- [47] R. S. Somai, M. J. Schut, and S. Van der Stigchel, "Evidence for the world as an external memory: A trade-off between internal and external visual memory storage," *Cortex*, vol. 122, pp. 108–114, 2020.
- [48] Stan Development Team, *RStan: The R Interface to Stan*, R package version 2.32.7, 2025.
- [49] B. W. Tatler and M. F. Land, "Vision and the representation of the surroundings in spatial memory," *Phil. Trans. Roy. Soc. B: Biol. Sci.*, vol. 366, no. 1564, pp. 596–610, 2011.
- [50] L. Talmy, "Lexicalization patterns: Semantic structure in lexical forms," in *Language Typology and Syntactic Description*, vol. 3, T. Shopen, Ed. Cambridge, UK: Cambridge Univ. Press, 1985, pp. 57–149.
- [51] A. Treisman and G. A. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, pp. 97–136, 1980.
- [52] F. J. Varela, E. Thompson, and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press, 1991.
- [53] A.-L. Vollmer, M. Mühlig, J. J. Steil, K. Pitsch, J. Fritsch, et al., "Robots show us how to teach them: Feedback from robots shapes tutoring behavior during action learning," *PLoS ONE*, vol. 9, no. 3, 2014.
- [54] P. C. Wason, "The contexts of plausible denial," *J. Verb. Learn. Verb. Behav.*, vol. 4, no. 1, pp. 7–11, 1965.
- [55] ELAN (Version 6.9) [Computer software]. Nijmegen, The Netherlands: Max Planck Inst. Psycholinguist., 2024.
- [56] M. Wilson, "Six views of embodied cognition," *Psychon. Bull. Rev.*, vol. 9, pp. 625–636, 2002.
- [57] J. K. Witt and M. A. Riley, "Discovering your inner Gibson: Reconciling action-specific and ecological approaches to perception-action," *Psychon. Bull. Rev.*, vol. 21, pp. 1353–1370, 2014.
- [58] R. A. Zwaan and G. A. Radvansky, "Situation models in language comprehension and memory," *Psychol. Bull.*, vol. 123, no. 2, pp. 162–185, 1998.



 Latest updates: <https://dl.acm.org/doi/10.1145/3686215.3689202>

RESEARCH-ARTICLE

Coupling of Task and Partner Model: Investigating the Intra-Individual Variability in Gaze during Human–Robot Explanatory Dialogue

AMIT SINGH, Paderborn University, Paderborn, Nordrhein-Westfalen, Germany

KATHARINA J ROHLFING, Paderborn University, Paderborn, Nordrhein-Westfalen, Germany

Open Access Support provided by:

Paderborn University



PDF Download
3686215.3689202.pdf
09 January 2026
Total Citations: 3
Total Downloads: 253

Published: 04 November 2024

Citation in BibTeX format

ICMI '24: INTERNATIONAL
CONFERENCE ON MULTIMODAL
INTERACTION

November 4 - 8, 2024
San Jose, Costa Rica

Conference Sponsors:
SIGCHI

Coupling of Task and Partner Model: Investigating the Intra-Individual Variability in Gaze during Human–Robot Explanatory Dialogue

Amit Singh

SFB/Transregio 318 Constructing Explainability
Paderborn University, Germany
amit.singh@uni-paderborn.de

Katharina J. Rohlfing

SFB/Transregio 318 Constructing Explainability
Paderborn University, Germany
katharina.rohlfing@uni-paderborn.de

Abstract

In a successful dialogue in general and a successful explanation in specific, partners need to account for both, the task model (what is relevant for the task) and the partner model (what one can contribute). The phenomenon of coupling between task and the partner model becomes especially interesting in the context of Human–Robot Interaction where humans have to deal with unknown capabilities of the robot, which can momentarily be perceived when the robot is unable to contribute to the task. Following research on the path over manner prominence in an action [31–33], a robot explained actions to a human by emphasizing two aspects – the path ("where" component) and the manner ("how" component). On critical trials, the robot occasionally omitted one of these components where participants sought missing information for the path or the manner. Participants' information-seeking and gaze behaviour were analysed. Analysis confirms the initial predictions for, a) task model (path over manner prominence), i.e., earlier information-seeking for path-missing than manner-missing trials, and b) partner model, i.e., while information-seeking is predominantly tied to the attention on the robot's face, when robot fails to provide resolution, attention shifts more often towards its torso – a behavior likely to indicate an exploration of the robot's capabilities. An individual-level analysis further confirms that the intra-individual variation in the task model is partly influenced by the perceived capability of the robot.

CCS Concepts

• Human-centered computing → Empirical studies in collaborative and social computing.

Keywords

Explanation, Scaffolding, Eyetracking, Partner Model, HRI

ACM Reference Format:

Amit Singh and Katharina J. Rohlfing. 2024. Coupling of Task and Partner Model: Investigating the Intra-Individual Variability in Gaze during Human–Robot Explanatory Dialogue. In *INTERNATIONAL CONFERENCE*

ON MULTIMODAL INTERACTION (ICMI Companion '24), November 04–08, 2024, San Jose, Costa Rica. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3686215.3689202>

1 Introduction

Explanation is a process as well as an outcome, which does not conform to a one-size-fits-all approach [30]. In this process, the involved partners dispose of distinct expectations about their partner (explainer) and the task [9, 18, 25]. Whereas the expectation about the partner relates to the questions of who is making what kind of contribution (partner model), the expectation about the task concerns which aspect of the task is relevant to achieve the task goal (task model) [37]. Although the coupling of the two models was postulated for effective tutorial interactions [37], little is known about how it can be achieved in other forms of dialogue, e.g., in Human–Robot Interaction. The coupling of such models becomes particularly significant in a human–robot explanatory dialogue, since in this context, humans as explainees not only question their partner's capabilities more often but also attempt to continuously determine whether and to what extent the robot can contribute to the task. For this purpose, humans usually test the robot with multiple probes on which basis they finally establish a partner model, i.e., what kind of capabilities they can rely on in an interaction [27]. Following this, the task model can be co-constructed during the course of interaction, tailored for both the parties based on their current state on understanding [25]. For example, studies in the past suggested that humans adapt to the robot's cues that are interpretative of what the robot is capable of [8]. In this sense, the dialogue "provides subtle clues to the robot's functionality and thus to adequate partner modeling" [8, p. 35]. Other studies have shown that humans are able to align with the robot's actions depending on their partner model of the robot [36]. However, previous studies did not model the individual differences to study how the coupling of these models can be achieved. For an individual, the task conceptualization could rely on the partner model, such that participants can structure or adapt their linguistic behavior about the task based on the perceived capability of the robot [35]. In this context, the perceived capability at a specific moment in the interaction can influence how the utterances about the source of misunderstanding are structured [25]. Whereas most studies consider one of the aspects, the challenge lies in systematically define and relate the task and the partner models. Hence, here our aim was to first lay down the definitions of these models in scope of our study and subsequently investigate the relationship between them.



This work is licensed under a Creative Commons Attribution International 4.0 License.

ICMI Companion '24, November 04–08, 2024, San Jose, Costa Rica
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0463-5/24/11
<https://doi.org/10.1145/3686215.3689202>

1.0.1 Task model. We define a task model requiring information on two dimensions to perform the given action – Path and Manner. Here, the path characterizes the “where” component in an action, whereas the manner specifies “how” the action is to be performed. This task model is grounded on an established finding in psycholinguistics research [29, 31], where each action was performed with attention to the path and manner. The explainer (robot) provided the information on how and where the given object has to be placed through a verbal guidance (section 2.4). In numerous studies, it has been found that in an action event while path remains in the foreground of attention the manner component tends to be attenuated, resulting in a more robust memory for the path compared to the manner [31–33]. This phenomenon is consistently observed across studies in adults as well as developmental studies in infants [14, 22]. Thus, we expected to observe the effect of path vs manner prominence in the information-seeking behaviour of the participants, consequently informing us about the task conceptualization. This approach first allows us to establish our initial predictions for the task model, and then systematically investigate how the partner model influences it during an interaction.

1.0.2 Partner model. We view the partner model as a dynamic process, evolving as the interaction progresses. In our study, the robot predominantly serves as the explainer, given that it possesses more information about the task than the explainee. Consequently, we assume that its capability is continually assessed by the participants (explainee) throughout the interaction, and it may be perceived as less capable when it fails to provide resolution occasionally. Therefore, the explainee develops an expectation or “need” for the robot to effectively scaffold the task. The partner model thus represents the explainee’s perception of the explainer’s (robot) capability to act as a scaffolding agent. Crucially, this initial perception of the explainer can change throughout the interaction based on the explainer’s demonstrated ability to provide – or fail to provide – adequate explanations when requested. Hence, the partner model predominantly draws inspiration from the concept of scaffolding, which is characterized as a supportive and dynamic form of assistance provided by a more knowledgeable partner, tailored to the learner’s abilities [37]. In HRI, this emphasizes that the explainer (robot) must be perceived as a partner possessing sufficient characteristics of a scaffolding agent before the explainee can rely on this support. In situations where the conversation breaks down due to a lack of sufficient information or misunderstanding, to keep the loop going, the explainee should be able to offload the task demands to the explainer (robot) by assigning the missing-information in the task. This assignment could manifest in rather various ways, such as looking at the robot’s face when in doubt or before raising requests [2, 7, 34], which is considered as a proactive function of gaze during information-seeking in conversation [17, 24].

1.1 Present Study

The setting was created by incorporating the abovementioned two models. In each trial, the robot asked the participant referring to an object to place in a specific manner (vertically or horizontally) on a destined location (path; on letters). To elicit questions from the participant, the robot occasionally omitted either the path or manner information. Our predictions were based on population

as well as individual level variations for both the task and partner models, which we specify hereunder.

1.2 Predictions

1.2.1 Task Model: Building upon the findings in psycholinguistics that the path remains salient than manner in an action [21, 29, 31], we predicted that the absence of path information would elicit quicker information-seeking behaviour from participants compared to the instances where the manner information is missing. Crucially, this prediction serves as a validation of our task model. The prediction is also supported by the previously proposed notion that the asymmetrical conceptualization of the path and manner arises from the intrinsic goal-directed nature of cognition, with a primary emphasis on the path [21, 32, 33].

1.2.2 Partner Model: Perceiving a partner as a legitimate scaffolding agent would entail focusing more on the partner’s face than other parts of the body especially while information-seeking phase. Crucially, this prediction would be supplemented by comparing the gaze behavior in the condition where the robot successfully provides a resolution (successful scaffolding), compared to a condition where the robot fails to provide a resolution (unsuccessful scaffolding), prompting the explainee to re-ask the question or proceed without support. We further predicted that an unsuccessful scaffolding would lead to an exploration of the partner capabilities where the attention shifts from the robot’s face to the other parts of the body. Moreover, evidence from human–human conversation suggests that not looking into the face while information-seeking is often interpreted as a sign of avoidance and lack of interest in conversation [11, 16], such as in the case of unsuccessful scaffolding. Thus, at the population level, we predicted that while the overall looks to face would be higher than other parts of the body during information-seeking phase, this facial engagement would decrease when the robot shows a lack of capability to scaffold the task.

1.2.3 Relationship between the task and partner model. At the individual level, we hypothesized that participants showing more exploratory behaviour while information-seeking – a behavior that is likely to reflect an exploration of partner’s capabilities following unsuccessful scaffolding – would exhibit more delay in information-seeking. Here our aim is to find out whether the source of the variability in the task model (delay in information-seeking) is partly determined by the uncertainty about the partner’s role and its potential scaffolding capabilities.

2 Method

The methods in this study are approved by the Review Board of the university and the informed consent from all the participants were obtained. The study was preregistered on Open Science Framework (OSF) before data acquisition, and the associated analysis code is available online and the data would be available on request.

2.1 Participants

Thirty-three university students (mean age = 22.30, SD = 2.63) were recruited via classroom advertisements and flyer distribution. All participants demonstrated native to fluent proficiency in German. Data from 3 participants could not be analysed due to track-loss



Figure 1: Setup and example stimulus: manners (vertical and horizontal) and paths (A N P I R); Three screws needed to be vertical, while the screw on R needed to be horizontal.

of eye-tracking sample points, hence the final sample size remains thirty participants.

2.2 Stimuli

The stimuli consisted of four objects in the shape of screw of different colours, Green, Red, Yellow, Blue (Fig. 1). Each object can be placed in two possible manners (we will refer to it as a "How"-component) vertically or horizontally (referred to as "hochkant" or "quer" in German). For complexity consideration, we limited the options to these two manner-affordances. Additionally, for each of the four objects, there were five potential positions marked on the table by unique alphabets where the object could be placed. These positions served as the path (we will refer to it as a "Where"-component). The combination of manners (vertical or horizontal) and paths (identified by unique alphabets) was generated randomly for each trial, resulting in a diverse array of configurations, offering distinct patterns concerning both manner and paths and maintaining variability in the task at the same time.

2.3 Procedure

As illustrated in the Figure 1, participants were seated in front of the NAO robot while wearing Pupil Labs glasses for eye-tracking (120 Hz), with a 5-point calibration yielding an accuracy of 0.60° and precision of 0.02° . The room was controlled for illumination and noise isolation. The objects were placed in front of the participant in a transparent glass container on a table. Alphabets stickers were pasted in a row to act as destination path of the object. Two stationary cameras were placed on the sides to capture the experimental setup. To ensure an unbiased interaction between the robot and the participant, an experimenter zone was established behind the participant, concealed by a divider. This setup ensured that the experimenter remained out of sight, facilitating uninterrupted interaction solely between the robot and the participant. The experiment started by calibrating the eyetracking and then introducing the participant to the task.

2.4 Task Instructions

Each participant was instructed to follow the robot for guidance on each placement of the object and were told to ask questions when they feel that the information for the task is insufficient. Since the missing information was only regarding the path or the manner, we expected the participants to ask questions along those dimensions (e.g., where, how, and so on). However, participants were given

the liberty to request a repetition whenever they feel it necessary, enhancing the flexibility of the interaction. After introducing the participants to the experiment, the experimenter retreated to the experimenter zone and initiated the experiment script. The robot was programmed with semiautonomous capabilities, where the control can be taken over by the experimenter in case the robot fails or conversation breaks down. On instances when the robot was unable to hear or respond, participants repeated the commands. The robot first introduced itself and then reiterated the task instructions again. After asking for the participant's readiness, the robot commences with the first trial. Each trial began with a specific instruction for placing an object in a designated manner and on a specific path. The instruction template was structured as follows:

*Place the green Object [Manner] on [Path]
Lege bitte das grüne Objekt [Hochkant or Quer] auf [A]*

Following each instruction, the robot gazed briefly to the task and then to the participant and sought confirmation: "Hast du das erledigt?" ("Have you done it? in English). Upon receiving affirmation, the robot progressed to the next object. The trial concluded when all four objects were placed on the designated path with specific manner. Subsequently, the robot asked participants to put back the objects and then turn back to solve a simple addition or subtraction task. This was done to offload the working memory and to keep participants engaged in the task rather than just passively following it. After solving the task, participants turned back towards the robot, and completed a recall task where they placed back the objects as instructed during the trial. Upon confirmation, the robot proceeded to the next trial. The experiment had four blocks, two for path and two for manner critical trials. Thus, a block included total five trials, and each trial had 4 moves corresponding to 4 objects, resulting in a total of 20 moves in a block generated randomly. In each block, information regarding the path or manner was intentionally omitted (missing information trials) randomly on 4 moves and rest of the moves (16) were of full information (full information trials). This was done to prompt questions from the participants for the missing information occasionally. The entire experimental session lasted approximately 45 minutes, during which eye-gaze and video data were continuously recorded.

3 Data Analysis

Eyetracking data was exported using the Pupil Player software. The videos were initially coded using ELAN [1] and later exported for further analysis. Since we were only interested in the behaviour for the time window from instruction onset till the moment of asking the question, we focus our analysis on this time window. All preprocessing and analysis were conducted using an R script [23], which is openly accessible in an online repository for reproducibility.

3.1 Preprocessing gaze data

Since we were interested in the looks to the face, torso, and the objects, three rectangular area of interests, referred to as surfaces, were defined around the robot's face, torso and the objects respectively. Fixations were computed using a time range from 80 to 350 ms with

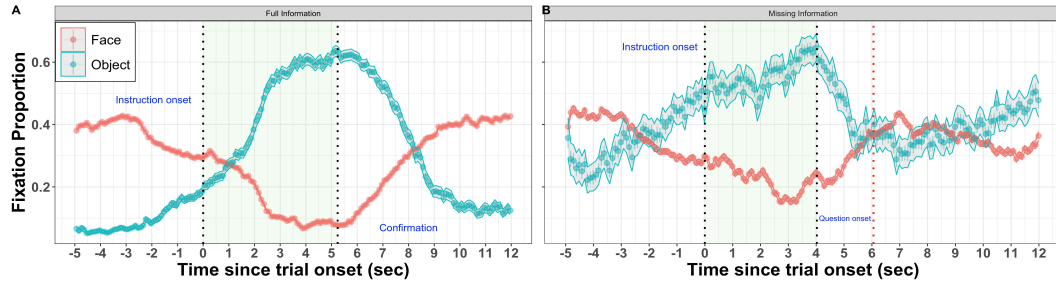


Figure 2: Gaze on robot's face and objects; (A) full information trial and, (B) missing path or manner information trial (right). The green-shaded region shows the instruction duration. Vertical red-line in B depicts the average question onset time

a maximum dispersion set at the default value of 1.5 degrees. To calculate the fixation proportion on a surface within a specified time range, the entire time window was divided at a 50 ms resolution. The number of fixations falling within a surface was then divided by the total number of fixations within that time bin. The delay in the information-seeking was calculated for the time window after the instruction offset by the robot for each trial. Our aim was to model the time-dependent changes in the gaze for the full and missing information trials. The full-information trial spanned from start of the instruction till the confirmation was provided by the participant. And the missing information trial spanned from start of the instruction till the offset of the participants' question. We modeled these behaviours on a continuous time scale using 2nd order Growth Curve Analysis (GCA), a methodology previously applied in analyzing time-dependent changes [19].

3.2 Preprocessing ELAN data

The video data was manual coded on ELAN by three student assistants independently, who were unaware of the hypothesis of the study. The codings included the annotations of robot instructions, participant-posed questions and successful vs unsuccessful scaffolding. A successful scaffolding was considered when the participant sought information and the robot responded successfully (No-error trials), otherwise an unsuccessful scaffolding when the robot did not respond to the request (error trials).

4 Results

4.1 Attention on the objects and the face

For the full-information condition (Fig. 2. A), we observed an increase of attention to the objects as soon as the instruction was presented (vertical line at 0). This continued throughout the instruction period (green shaded region), where a transient decrease in fixation on the face was observed, showing the participants' anticipatory fixation to the objects upon encountering the instruction. Crucially, the attention to objects remained till the end of instruction, after which there was a shift towards the face, as a result of providing confirmation for the completed task (around 10 sec). For the missing information trial (Fig. 2. B), there was an increase in attention towards the objects as soon as the instruction was encountered (vertical line at 0). However, at the end of the instruction, and after realizing that the information was insufficient, participants quickly redirected their gaze back to the robot's face, likely to seek information for the missing aspect in the task. This information-seeking is shown by a transient increase in gaze on robot's face starting just after the offset of instruction which suggests that information-seeking is tightly coupled with looking at the partner, and in this case specifically on the face. This increase in looks to the face was followed by an information-seeking (red vertical line). Importantly, the delay in the onset of this information-seeking was depended upon the type of information missing in the task, i.e., the path or the manner (task-model). According to our first prediction we expected this delay to be higher for the manner than the path, which we analysed subsequently.

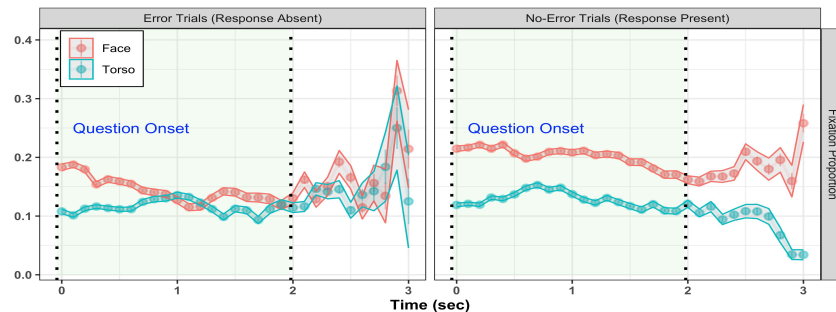


Figure 3: Fixation on face and torso for critical trials where robot did or did not respond to the question. Left plot; unsuccessful scaffolding, Right plot; successful scaffolding. The shaded region shows the mean duration of question onset and offset.

4.2 Information-seeking delay for path and manner

Following our prediction about the task model, we compared the time of question onset for the path and manner critical trials. Here, a delayed onset was observed for manner-related questions compared to path-related questions (Fig. 4), supporting our initial prediction regarding path prominence in event conceptualization. This pre-

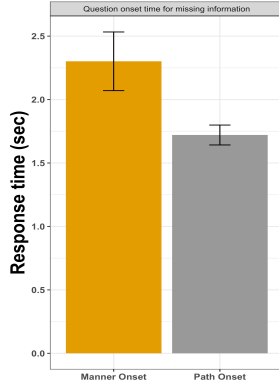


Figure 4: Difference in path vs manner question time. Conditions are on x-axis and response time on y-axis

diction was confirmed by fitting a hierarchical linear mixed effects model using lme4 package in R [5]. We incorporated a full model with varying intercepts and slopes for each participant to account for main and random effect structure. The predictors were sum contrast coded, with manner coded as +0.5 and path as -0.5, such that the intercept represented the overall mean delay, and the slopes indicated deviations around this mean for each condition. Model predictions are depicted in Fig. 4. Confirming our predictions for the task model, we found a significant delay for the manner than the path related questions ($\beta = 0.44$, S.E. = 0.2, $p=0.03$).

4.3 Task and partner model

Our population-level partner model suggested that participants redirect their attention towards the partner's face from the object when they intend to ask question (Fig. 2). Following our partner model predictions, whether the participants' perception about the robot's scaffolding capability (successful vs unsuccessful scaffolding) guided the gaze behaviour towards the robot, we analyzed the gaze in two conditions – where robot successfully provided the resolution (Response present) or did not provide the resolution (Response absent) (Fig. 3). Crucially, we looked into the fixation on the robot's face and the torso during the information-seeking phase. For the unsuccessful scaffolding, we observed an exploratory gaze behaviour where a higher gaze switches from robot's face to the torso and vice-versa was observed (Fig. 3, Response absent or error trial). Importantly, when robot successfully scaffolded the task (Response present), the attention was predominantly was on the face throughout the entire information-seeking phase (Fig. 3, Response present or No-error trial). Further to investigate whether the

scaffolding capability contributes to the task conceptualization and thus influence the task model (information-seeking), we modeled the attention allocation on face vs torso as a function of question onset delay. The aim was to look whether the response delay predicted the individual gaze preferences on the face over the torso. For this, we focused our analysis in the information-seeking time window, since this specific window allowed us to examine the proactive role of gaze, when one seeks information particularly about the task. A linear mixed effect model was fitted for logit transformed proportion of looks at the face and torso following [3]. The model included sum-contrast coded Area of Interest (AOI: Face vs Torso) and the question onset delay as a continuous predictor by mean scaling. Our goal was to fit a maximal model by accounting for random effects at the subject level variation with slope correlations[4]¹. Hence, the intercept represented the grand mean of gaze towards the face and torso for mean question onset delay, and deviation around this mean as the effect size of looks toward the face as compared to the torso for the population and the individual subjects. Importantly, we were interested in the random effects (correlation parameter) between the gaze and delay effect sizes for each subject. A full and reduced model comparison was done using log-likelihood test by removing only the correlation parameter from the model [4]. The model predictions for the subjects are depicted in Figure 5. Adding

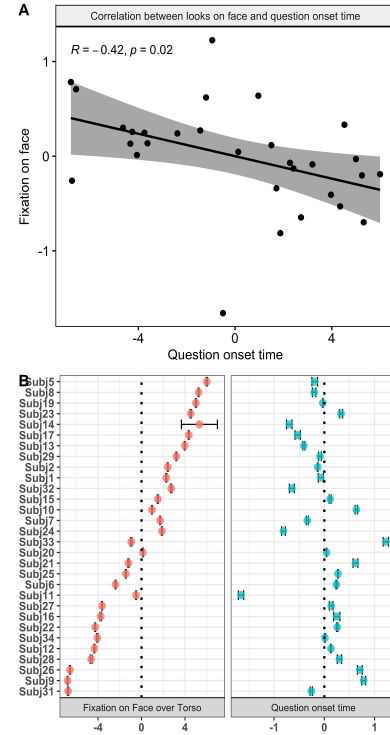


Figure 5: (A) Correlation, looks on face over torso and question time, (B) model prediction for participants

¹model: lmer(meanLook <- 1+AOI*CentMeanDelay + (1+AOI+CentMeanDelay|Subj))

the correlation random structure between gaze and delay slope parameters significantly improved the model fit ($\chi^2 = 997.86$, $Df = 1$, $p < 0.001$), suggesting that the correlational parameter effectively captured the variation between onset delay and the gaze on face over torso. Further, there was a significant interaction between AOI and question onset time ($\beta = -0.14$, $SE = 0.0$, $p < 0.001$), such that the delay also negatively predicted the gaze on the face over the torso. Meaning, subjects showing more preference for the robot's face than the torso were quicker in asking questions ($R = -0.42$, $p = 0.02$), as depicted in Figure 5. In comparison to our population-level partner model, where the majority of participants attended to the face while asking question, on the contrary individuals who exhibited a higher tendency to look at the torso over the face specifically during information-seeking phase, experienced more delay in information-seeking for critical trials. To highlight again, more preference to the torso was observed for the condition where robot was perceived less capable of scaffolding the task (Fig. 3)

5 Discussion

The study builds upon a task model, that requires a differentiation between path and manner of an action. Further, the partner model consists of the expectation that the partner can scaffold the participant to perform those actions. Apart from establishing the validity of these models at the group level, our focus was predominantly to account for the intra-individual variability of the task and the partner model in HRI. Based on our initial assumptions about the task model, we observed that participants were generally quicker in information-seeking for the path compared to the manner. This aligns with a robust finding in psycholinguistics [6, 20–22], which we tested in the context of HRI.

Our investigation into the partner model, with a focus on attentional dynamics, confirmed our initial assumption. At the population level, participants tended to look more at the robot's face during two distinct time segments: a) while receiving the instructions and b) while information-seeking (asking questions). Crucially, while receiving the instructions, we also observed a transient decline in looks at the face which potentially shows the anticipatory attention to the objects as the task instruction is unveiled. Conversely, when participants asked questions (for missing information), their gaze was predominantly directed towards the robot's face—a behavior indicative of perceiving the robot as a scaffolding partner. It is worth to mention that despite the fact that area of interest for the body covered the most part of the visual field of view, the participants' gaze on the face was predominantly higher which might reflect that the robot was ascribed as a scaffolding partner over and above just a point of verbal reference. In this case, one could have sought information by keeping attention on the objects, which might be conceived as a question to oneself than to the partner, given the proactive role of gaze in day-to-day social encounters [11, 15].

Moreover, we observed a delay in manner than path related questions. This underlines our inherent task model, supporting the idea that manner carries more vulnerability necessitating conscious cognitive operations, as opposed to the relatively automatic processing of path relying on less cognitive resources [21]. As a definition, a conscious cognitive operation on a representation would require more scaffolding than a representation which is already in

attention (e.g., path). One of the methods suggested in previous studies to mitigate this effect is contrastive scaffolding of the path and manner [13, 28], which has been shown to improve attention to both components. Provided that in our experiment the path and manner instructions were presented randomly, and the delay was observed for linguistic encoding, than performing an action per se, the likely explanation of the manner related cost could be due to linguistic processes governing the retrieval of these two aspects.

Nevertheless, scaffolding the manner might need more intervention of the scaffolding partner than the path. In this regard, the intra-individual findings from the study highlight the aforementioned aspects, quicker information-seeking when the preference to the robot's face is higher than the torso. This might suggest that the delay in linguistic processing was partly influenced by whether the robot was conceptualized as a partner possessing sufficient scaffolding capability or not. Especially, when the robot did not successfully scaffold the task, the gaze was more exploratory where the attention shifted more often to the torso and other body parts, indicating an exploration.

Here, we focused on a specific aspect of the partner model – perceived scaffolding capability – and found that participants' delayed response is partly influenced by their perception of the partner's scaffolding capability (as assessed by the gaze). We should critically note that in future, our assessment needs to be supported by results from either an online measure assessing the dynamical change in perception of partner's role or an offline survey. Our findings bear significance not only in the realm of human–robot interaction but also extend to human–human interactional contexts, i.e., perceiving a partner capable of scaffolding might allow the explainee to flexibly ascribe more responsibility for the task to their partner. This is grounded on the understanding that joint actions involve close collaboration and a shared distribution of the task load among participants [10]. For example, the amount of information an explainee shares about the source of a misunderstanding can offer insights into her commitment to a joint goal. Furthermore, when an explainee use varied constructions in their verbal behaviour, it can also establish the partner model as more capable to scaffold the task as opposed to when they use repetitive requests, since day-to-day language use is hardly deterministic and relies on variability [12]. In future, we aim to capture more qualitative aspects of this perceptual dimension and investigate its meaning for the facial engagement. Other functional aspects of these measures should also not be neglected, which might entirely allude to a different cognitive function in an explanation, such as, monitoring, which has been proposed to play a different role in explanation processes [26]. Including a more holistic scale in such cases additionally requires considering the meta aspects from multimodal interaction (e.g., backchanneling and signs of ignoring the partner) and might offer us a deeper insight about explanation processes in HRI.

6 Conclusion

This study suggests a potential pathway to systematize task and partner models from an interdisciplinary perspective by investigating human Gaze behaviors in HRI. Future research should delve into the proactive role of the partner model in shaping task model, particularly when they are embedded in an interactional context.

Acknowledgments

The work is supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824.

References

- [1] 2023. *ELAN (Version 6.7)*. <https://archive.mpi.nl/tla/elan> [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive..
- [2] Michael Argyle and Roger Ingham. 1972. Gaze, Mutual Gaze, and Proximity. *Semiotica* 6, 1 (1972), 32–49. <https://doi.org/10.1515/semi.1972.6.1.32>
- [3] Dale Barr. 2008. Analyzing ‘visual world’ eyetracking data using multilevel logistic regression. *Journal of Memory and Language* 59, 4 (2008), 457–474. <https://doi.org/10.1016/j.jml.2007.09.002>
- [4] Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68, 3 (2013), 28–38. <https://doi.org/10.1016/j.jml.2012.11.001>
- [5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- [6] Philippe Bourdin. 1997. On Goal-Bias Across Languages: Modal, Configurational and Orientational Parameters. In *Typology: Prototypes, Item Orderings and Universals (Proceedings of LP '96)*, Bohumil Palek (Ed.). Charles University Press, Praha, 185–218.
- [7] Michael Cook. 1977. Gaze and mutual gaze in social encounters: How long—and when—we look others “in the eye” is one of the main signals in nonverbal communication. *American Scientist* 65, 3 (1977), 328–333.
- [8] Kerstin Fischer. 2011. How people talk with robots: Designing dialog to reduce user uncertainty. *AI Magazine* 32, 4 (2011), 31–38.
- [9] Alan Garfinkel. 1982. Forms of Explanation: Rethinking the Questions in Social Theory. *British Journal for the Philosophy of Science* 33, 4 (1982), 438–441.
- [10] Margaret Gilbert. 2014. *Joint Commitment: How We Make the Social World*. Oxford University Press.
- [11] Erving Goffman. 1964. The Neglected Situation. *American Anthropologist* 66, 6 (December 1964), 133–136. <https://doi.org/10.1525/aa.1964.66.suppl.3.02a00030>
- [12] H. P. Grice. 1971. Intention and Uncertainty. *Proceedings of the British Academy* 57 (1971), 263–279.
- [13] André Groß, Amit Singh, Ngoc Chi Banh, Birte Richter, Ingrid Scharlau, Katharina J. Rohlfing, and Britta Wrede. 2023. Scaffolding the Human Partner by Contrastive Guidance in an Explanatory Human-Robot Dialogue. *Frontiers in Robotics and AI* 10 (2023), 1236184. <https://doi.org/10.3389/frobt.2023.1236184>
- [14] Susan J. Hespos, Molly M. Saylor, and Sharon R. Grossman. 2009. Infants’ Ability to Parse Continuous Actions. *Developmental Psychology* 45, 2 (2009).
- [15] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26 (1967), 22–63. [https://doi.org/10.1016/0001-6918\(67\)90005-4](https://doi.org/10.1016/0001-6918(67)90005-4)
- [16] Robin H. Kendrick and Judith Holler. 2017. Gaze Direction Signals Response Preference in Conversation. *Research on Language and Social Interaction* 50, 1 (2017), 12–32. <https://doi.org/10.1080/08351813.2017.1262120>
- [17] Christopher L. Kleinke. 1986. Gaze and Eye Contact: A Research Review. *Psychological Bulletin* 100, 1 (1986), 78–100. <https://doi.org/10.1037/0033-2909.100.1.78>
- [18] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (2019), 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- [19] D. Mirman, J.A. Dixon, and J.S. Magnuson. 2008. Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language* 59, 4 (2008), 475–494.
- [20] Anna Papafragou. 2010. Source-Goal Asymmetries in Motion Representation: Implications for Language Production and Comprehension. *Cognitive Science* 34, 6 (August 2010), 1064–1092. <https://doi.org/10.1111/j.1551-6709.2010.01107.x>
- [21] Sarah Pourcel. 2004. What makes path of motion salient?. In *Proceedings of the 30th Annual Meeting of the Berkeley Linguistics Society*. Sheriden Books, Ann Arbor, MI, 505–516.
- [22] Shannon M Pruden, Tilbe Gökşun, Sarah Roseberry, Kathy Hirsh-Pasek, and Roberta M Golinkoff. 2012. Find your manners: How do infants detect the invariant manner of motion in dynamic events? *Child Development* 83, 3 (2012), 977–991. <https://doi.org/10.1111/j.1467-8624.2012.01741.x>
- [23] R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- [24] F. Roassano. 2010. Questioning and responding in Italian. *Journal of Pragmatics* 42, 10 (2010), 2756–2771.
- [25] Katharina J. Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M. Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, Ilona Horwath, Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel-Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. 2021. Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. *IEEE Transactions on Cognitive and Developmental Systems* 13, 3 (Sept. 2021), 717–728. <https://doi.org/10.1109/TCDS.2020.3044366>
- [26] Katharina J. Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M. Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, Ilona Horwath, Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel-Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. 2021. Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. *IEEE Transactions on Cognitive and Developmental Systems* 13, 3 (2021), 717–728. <https://doi.org/10.1109/TCDS.2020.3044366>
- [27] J. Scholtz. 2003. Theory and evaluation of human robot interactions. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*. 10 pp.–. <https://doi.org/10.1109/HICSS.2003.1174284>
- [28] Amit Singh and Katharina Rohlfing. 2023. Contrastiveness in the Context of Action Demonstration: An Eye-Tracking Study on Its Effects on Action Perception and Action Recall. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 45. <https://escholarship.org/uc/item/2w94t4cv>
- [29] Dan I. Slobin. 1987. Thinking for Speaking. In *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, Vol. 13. 435–445. <https://doi.org/10.3765/bls.v13i0.1826>
- [30] Karol Sokol and Peter Flach. 2020. One Explanation Does Not Fit All. *Künstliche Intelligenz* 34 (2020), 235–250. <https://doi.org/10.1007/s13218-020-00637-y>
- [31] L. Talmy. 1975. Semantics and syntax of motion. In *Syntax and Semantics*, J. Kimball (Ed.). Vol. 4. New York Academic Press, New York.
- [32] Leonard Talmy. 1985. *Lexicalization patterns: Semantic structure in lexical forms*. <https://www.degruyter.com/database/COGIB/entry/cogib.11708/html?lang=en>
- [33] Leonard Talmy. 2000. *Toward a Cognitive Semantics*. 44, Vol. II. <https://mitpress.mit.edu/9780262700962/toward-a-cognitive-semantics/>
- [34] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. 2001. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) (CHI '01). Association for Computing Machinery, New York, NY, USA, 301–308. <https://doi.org/10.1145/365024.365119>
- [35] A.-L. Vollmer, K.S. Lohan, K. Fischer, Y. Nagai, K. Pitsch, J. Fritsch, K. Rohlfing, and B. Wrede. 2009. Early Information-Seeking in Human-Robot Interaction: Face vs. Torso Attention Shifts. In *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on Development and Learning*. IEEE, 1–6. <https://doi.org/10.1109/DEVLRN.2009.5175525>
- [36] A. L. Vollmer, K. J. Rohlfing, B. Wrede, and A. Cangelosi. 2015. Alignment to the Actions of a Robot. *International Journal of Social Robotics* 7, 2 (2015), 241–252.
- [37] D. Wood, J. S. Bruner, and G. Ross. 1976. The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry* 17, 2 (1976), 89–100.

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

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

Permalink

<https://escholarship.org/uc/item/2w94t4cv>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors

Singh, Amit

Rohlfing, Katharina

Publication Date

2023

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

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 sub-actions 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

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 de-

mands 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; Bungler, 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!].

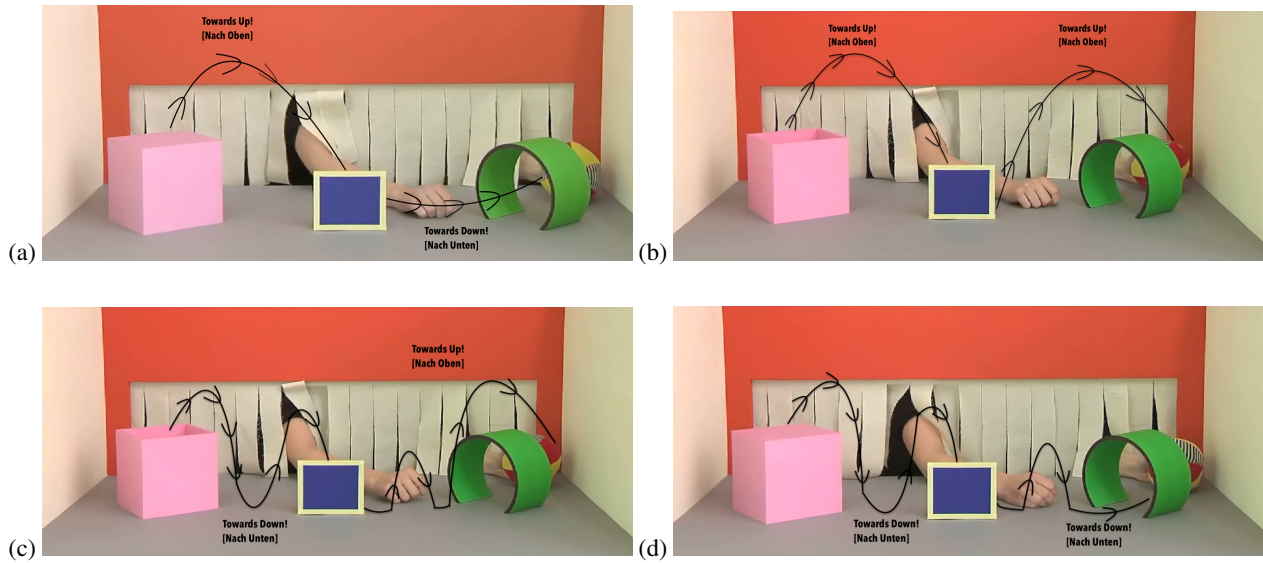


Figure 1: Stimulus videos depicting the path conditions; (a) **contrastive** [Up-Down], (b) **non-contrastive** [Up-Up], (c) **contrastive** [Down-Up], (d) **non-contrastive** [Down-Down]. Black lines show the movement path of the ball with instruction in **assertion-assertion** condition

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..

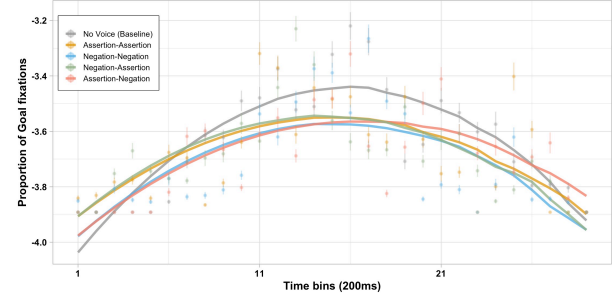


Figure 2: Proportion of Goal fixation for **Contrastive** motion

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.	<i>t</i>	<i>p</i>
(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. Import-

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.

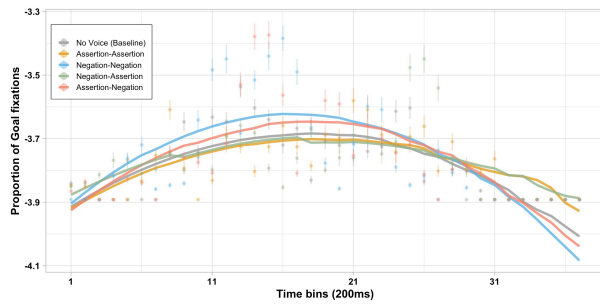


Figure 3: Proportion of Goal fixation for **non-contrastive** motion

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).

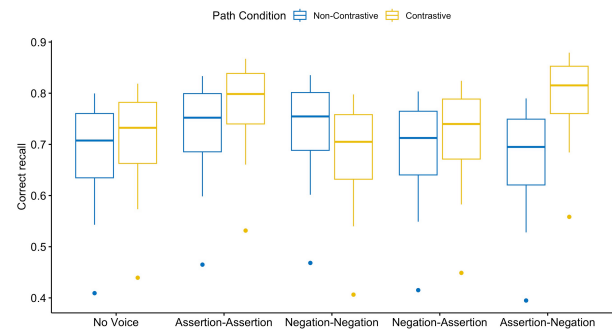


Figure 4: Proportion of correct recall for contrastive and non-contrastive path conditions for instruction conditions plotted against x-axis

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 ($\beta = 0.47$, $SE = 0.16$, $t = 2.83$, $p < 0.00$) than assertion–assertion condition ($\beta = 0.37$, $SE = 0.16$, $t = 2.23$, $p < 0.00$), 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	<i>S.E.</i>	<i>t</i>	<i>p</i>
(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	<i>S.E.</i>	<i>t</i>	<i>p</i>
(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	<i>S.E.</i>	<i>t</i>	<i>p</i>
(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*, 4(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.

Journal of Psycholinguistic Research

A methodological framework for studying online language–vision interplay in human–robot interaction

--Manuscript Draft--

Manuscript Number:	JOPR-D-26-00014R1	
Full Title:	A methodological framework for studying online language–vision interplay in human–robot interaction	
Article Type:	Original Research	
Keywords:	Eye movements; human--robot interaction (HRI); repairs; interactive methods	
Corresponding Author:	Amit Singh, Ph. D. Paderborn University Faculty of Arts and Humanities: Universitat Paderborn Fakultat fur Kulturwissenschaften GERMANY	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	Paderborn University Faculty of Arts and Humanities: Universitat Paderborn Fakultat fur Kulturwissenschaften	
Corresponding Author's Secondary Institution:		
First Author:	Amit Singh, Ph. D.	
First Author Secondary Information:		
Order of Authors:	Amit Singh, Ph. D.	
	Katharina J. Rohlfing	
Order of Authors Secondary Information:		
Funding Information:	Deutsche Forschungsgemeinschaft (2021-438445824)	Prof. Dr. Katharina J. Rohlfing
Abstract:	This article presents a methodological framework for studying online language--vision processes in an interactive HRI setup. In such a setting, language-mediated eye movements can be examined as dialogue unfolds between a human and a robot. Traditionally, real-time linguistic processes have been studied using visual world paradigms (VWP), in which either comprehension or production tasks are implemented on screens under tightly controlled conditions. Going beyond these traditional and largely unidirectional approaches, we use a humanoid robot to bring comprehension and production into a single interactive framework while retaining experimental structure. Using one proof-of-concept empirical example, we outline the methodological potential of HRI setups for investigating online language--vision interplay in dialogue. More specifically, we show how a traditional VWP design can be embedded within an interactive HRI task where the participant not only engages with a partner but at the same time acts on the physical stimuli. When such tasks are grounded in robust psycholinguistic assumptions and empirical predictions, they can serve as a controlled interactive test-bed for examining whether established findings from real-time language processes extend to richer task contexts. We illustrate this approach with one example involving action structure (path vs.\ manner) and repair format (open vs.\ restricted) to argue that HRI can offer a useful methodological middle ground for studying online language--vision processes in a setting that is more interactive than traditional screen-based paradigms.	

A method for studying online language–vision interplay in a human–robot interaction

Amit Singh and Katharina J. Rohlfing

Psycholinguistics, Faculty of Arts and Humanities, Paderborn University, Germany

Funding: This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021-438445824 “Constructing Explainability”

Consent and approval This study was approved by University ethics board, and performed in accordance with relevant guidelines and regulations. Informed consent was obtained from all individual participants included in the study.

Data availability The data supporting the findings of this study and the analysis code are available in an anonymized OSF repository (view-only link):

https://osf.io/nxzqs/?view_only=99585617fb144c76b2c44715e8bd1d7f

Conflict of interest: The authors declare no conflict of interest.

A methodological framework for studying online language–vision interplay in human–robot interaction

Abstract

This article presents a methodological framework for studying online language–vision processes in an interactive HRI setup. In such a setting, language-mediated eye movements can be examined as dialogue unfolds between a human and a robot. Traditionally, real-time linguistic processes have been studied using visual world paradigms (VWP), in which either comprehension or production tasks are implemented on screens under tightly controlled conditions (Tanenhaus et al., 1995). Going beyond these traditional and largely unidirectional approaches, we use a humanoid robot to bring comprehension and production into a single interactive framework while retaining experimental structure. Using one proof-of-concept empirical example, we outline the methodological potential of HRI setups for investigating online language–vision interplay in dialogue. More specifically, we show how a traditional VWP design can be embedded within an interactive HRI task where the participant not only engages with a partner but at the same time acts on the physical stimuli. When such tasks are grounded in robust psycholinguistic assumptions and empirical predictions, they can serve as a controlled interactive test-bed for examining whether established findings from real-time language processes extend to richer task contexts. We illustrate this approach with one example involving action structure (path vs. manner) and repair format (open vs. restricted) to argue that HRI can offer a useful methodological middle ground for studying online language–vision processes in a setting that is more interactive than traditional screen-based paradigms while remaining experimentally structured.

KEYWORDS

Eye movements, human–robot interaction (HRI), repairs, interactive methods

Introduction

For decades, psycholinguistics has grappled with a recurring tension between precise experimental control and the study of language in richer contexts (MacWhinney, 2000; Tanenhaus & Trueswell, 2006; Yule, 1996; Zwaan, 2016). This tension has given rise to two influential orientations along theoretical and methodological lines within psycholinguistics, namely “language-as-product” and “language-as-action” as dubbed by Clark (1992). Whereas the *language-as-product* orientation has examined language processing under tightly controlled conditions to investigate time-locked, real-time processes of comprehension or production, without goal-directed behavior (Altmann & Kamide, 1999; Chomsky, 1965; Jackendoff, 2002; Levelt, 1989; Tanenhaus et al., 1995), the *language-as-action* orientation has investigated language using offline methods under interactive setups with goal-directed and joint activity, where meaning is negotiated through social interaction in rich contexts (Clark, 1992, 1996; De Jaegher et al., 2010; Goodwin, 2000; Pickering & Garrod, 2004; Tomasello, 2003). In the past, there have been multiple attempts to bring these two orientations together for their complementary strengths, but they have not been able to capture typical assumptions of both *action* and *product* paradigms. For instance, an early attempt by Spivey et al. (2002) brought Visual World Paradigms (henceforth VWP) closer to an interactive approach. It preserved important aspects of a *language-as-action* situation, but did not fully realize interactivity. While Spivey et al. (2002) were able to extend the context from screens to real-world settings with participants having a clear behavioral goal, they nevertheless acknowledged that more work was needed to bring the two orientations together in VWP experiments. Subsequently, following works from Brown-Schmidt and Tanenhaus (2008) and Tanenhaus and Brown-Schmidt (2008) moved a step further to measure time-locked responses in unscripted conversation.

Whereas these early setups were able to capture essential aspects of *action* paradigms, such as spoken language as a medium of communication, a well-defined context, and a task with clear behavioral goals, they nevertheless did not fully implement interactivity, for two likely reasons, 1) scripted instructions as in Spivey et al. (2002), or 2) unscripted conversation, though within a partly restricted setup, e.g., curtain in-

between participants to preserve a substantial degree of control as in Brown-Schmidt and Tanenhaus (2008) and Tanenhaus and Brown-Schmidt (2008). While it is agreed that for a controlled yet interactive VWP setup appropriate trials and baselines have to emerge from the conversational interaction itself (which, ideally, should be mostly unscripted), the partner is nevertheless hidden to systematically separate *partner-driven* effects from *language-driven* eye movements. Introducing a human partner into the setup certainly involves a trade-off, as it comes at the cost of measurement precision for time-locked, language-driven eye movements, given that human partners introduce additional attentional shifts inherent to rich discourse and non-verbal behavior. Whereas introducing confederates to generate verbal instructions in the experiments is often cited as a way for control and reduce random variations, they nevertheless eliminate the natural collaborative process (Kuhlen & Brennan, 2013).

In this paper, we introduce a humanoid robot as a partner to mitigate sources of randomness on the partner’s side and to standardize partner behavior across different interactions. By doing so, the present article does not aim to resolve the broader theoretical differences between action paradigms and product approaches to studying human interaction. Rather, it opens up a new possibility to validate the already established rich theories in VWP on online language–vision interaction in HRI, while at the same time treating this new domain as its own field. It proposes a methodological middle ground in which *language-as-action* aspects are preserved through a bidirectional human–robot setup, while classical VWP methods are brought in to investigate language–vision interaction with the temporal precision traditionally associated with *language-as-product* paradigms in a more structured setting.

Since the inception of visual world paradigms (VWP) about 30 years ago (Tanenhaus et al., 1995), there is little doubt that eye-tracking has proven valuable for examining multifarious patterns of language–vision correspondence (Altmann & Kamide, 1999; Huettig et al., 2011; Tanenhaus & Trueswell, 2006). Historically, one of the main motivations to study online language processes was to demonstrate that language and vision are interactive, challenging the prevailing notion of modularity, and that visual context can influence speech processing and vice-versa (Anderson et al., 2011). In a controlled environment with visual stimuli, researchers were able to investigate when

and how referential expressions, e.g., *Pick up the Lamp* activate phonologically similar competitors over time (Allopenna et al., 1998). For example, a target word like *Lamp* may initially co-activate onset competitors like *Ladder* within first 200 ms, followed by a rhyme competitor, such as *Camp*, and subsequently by a semantically related item like *Table*. At the syntactic level, studies investigated how listeners handle temporary ambiguities in sentence processing (e.g., garden path sentences) and whether visual context facilitates such disambiguation (Spivey et al., 2002). Central to this kind of investigation was the aim to capture how linguistic ambiguity is resolved when visual context is made available (Spivey, 2007).

The question of referential resolution in *language-as-product* research is not fundamentally different from that in *language-as-action* settings; the main difference lies in the context in which it is investigated, namely communication tasks organized around joint goals (Brown-Schmidt & Tanenhaus, 2008; Clark & Wilkes-Gibbs, 1986; Spivey et al., 2002). Part of this difference in perspective lies in how the notion of “interaction” has been developed across the field of social cognition and classical cognitive science. Whereas in cognitive science there has been a call for a “paradigm drift” away from modularity (recent call, Spivey, 2023), social cognition emphasized distributed cognition and participatory aspects of social interaction at a broader level (recent call, Dingemanse et al., 2023). Rather than treating these approaches as methodologically different, it is more useful to view them as complementary, emphasizing different contexts for language investigation at both individual and interactional levels. For example, although traditional VWP experiments in psycholinguistics inform us about how listeners process referential expressions in real-time, they often do not inform us about interactive and communicative role of gaze in the presence of a partner (Goffman, 1959; Kingstone, 2020).

This methodological specificity further raises a broader question: how do the gaze patterns observed during real-time comprehension or production in tightly controlled psycholinguistic experiments generalize to the situations where comprehension and production is happening simultaneously in presence of an interaction partner? In fact, developmental literature already suggests that infants behave differently when a partner is present with whom they can create actions (Reddy et al., 2013). Recent work

also shows that partner-specific information, common and privileged ground can shape visual attention in real-time during language comprehension and production (Galati & Brennan, 2021; Wei et al., 2024). In contrast to an individual mind where the eyes do not signal information, but they solely serve to take information in, the presence of a partner requires the listener to not only passively gather information but also manage communicative signals by being aware of the partner and adjusting the looking behavior that reflects both comprehension and coordination (Kingstone, 2020; Richardson et al., 2007). In such settings, looking toward the partner can function as a cue that is relevant for coordination, while looking toward referential stimuli may reflect information acquisition mirroring the processes to those studied in VWP.

We see human-robot interaction (HRI) as a possibility for bringing the online experimental rigor of VWP to a very specific case of interaction, where human language processes are shaped moment by moment as they interact with a robot. Our proposal aims not only to advance research on online methodologies, but also to lay the groundwork for testing, validating and extending existing findings in an HRI setting. Although numerous studies have used HRI to investigate language processing, most have focused on outcome measures, such as interaction quality, anthropomorphism, personality traits and learning outcomes rather than the fine-grained analysis of real-time linguistic processes. To our knowledge, there have been very few experiments that have used robots to investigate online language processes in truly interactive settings. Moreover, those studies which have attempted to investigate language processes in an HRI setup, have examined either comprehension (e.g., Staudte & Crocker, 2009), or production processes in isolation (e.g., Stratou et al., 2021), and have not brought two processes within a unified framework.

Our call is also informed by the contemporary rapid advancements of artificial intelligence and large language models, which now offer myriad opportunities to tailor dialog models and create a situation where critical trials and baselines emerge as a process of interaction rather than scripted instructions. In this article, starting from a trial based experiment, we provide one proof-of-concept where a VWP experiment can be brought into HRI design to show that a real-time investigation of linguistic processes can be studied without abandoning interactive task structure.

The remainder of this paper is organized into four main sections followed by future directions. The first section contextualizes interactivity in psycholinguistic research. The second section explores how HRI can contribute to this domain by providing a more interactive yet structured setting for studying online language processes. The third section presents referential ambiguity resolution as a central methodological paradigm, one that has historically driven psycholinguistic inquiry and remains central in social interaction. The fourth section extends the scope of linguistic processing to action, illustrating how language–action coupling helps to achieve interactivity in an HRI-based setup through examples from our recent studies.

Toward interactivity in psycholinguistic research

In a recent article, Maselli et al. (2023) expressed optimism about the feasibility of conducting ecologically valid studies, particularly in light of recent technological advancements and emphasized that future research should harness these developments within more innovative experimental designs. Although the concept of interactivity in language study is by no means new, the way the multidisciplinary field of psycholinguistics has emerged and stood apart from these fields is related to an approach that has remained coherent in both method and theory (Levelt, 2013). Many studies in these related disciplines have relied on corpus analysis of existing texts and newspapers or on naturalistic data collected from phone conversations with strangers and video recordings “in the wild” (MacWhinney, 2000). Through such data, researchers have uncovered different patterns in communication, such as word frequency distribution (Piantadosi, 2014), lexical regularities (Dale & Spivey, 2006) and turn-taking timings (Stivers et al., 2009). However, these naturally occurring settings often present limitations for the types of phenomena that interest cognitive scientists, particularly those who were typically interested in linguistic processes taking place under constrained temporal and spatial conditions. For example, in controlled interaction where participants collaborate on a joint object assembly task, the phenomena of lexical alignment can be meaningfully investigated by tracking how participants’ use of terminology changes over time (Isaacs & Clark, 1987). In parallel with corpus-based and naturalistic approaches, in

a quest to investigate language in psychological terms, studies increasingly focused on uncovering the underlying cognitive processes by imposing strong temporal constraints. Tasks such as priming and lexical decision were employed to examine, for example, the time it takes to retrieve words from memory (D. E. Meyer & Schvaneveldt, 1971; Neely, 1977). The effect of context on the lexical accessibility was also tested, typically for a predefined set of sentential and lexical contexts, which revealed that context plays a critical role in modulating how and when words are accessed during comprehension (Stanovich & West, 1983). The story remained dominant until the visual context was also shown to concurrently affect multiple linguistic levels, including lexical, syntactic, and later even semantics and pragmatics processing (Knoeferle et al., 2005; Spivey, 2007; Spivey et al., 2002). These advances in studying temporal dynamics of language processing were made possible by technological advancements, most notably by the development of eye-trackers, where the VWP, in particular, would establish itself as a standard method in psycholinguistics to study online language comprehension (Huettig et al., 2011; Spivey, 2007; Tanenhaus et al., 1995). This technological shift marked a major stride forward by allowing the researchers to move beyond investigating language “in the wild”, toward investigating linguistic processes in semi-naturalistic and ecologically richer situations within an experimentally controlled environment by bringing the visual contexts into the lab. Hence, in retrospect, the convergence of technological advancements and innovative experimental design can be seen as a key driving force behind the rapid progress in psycholinguistics (Maselli et al., 2023; Ryskin & Spivey, 2023). By preserving visual context within structured tasks, VWP-based experiments offered a compromise between ecological validity and methodological rigor, though not without certain limitations (McMurray, 2023; Ryskin & Spivey, 2023). Here, ecological validity is treated as a matter of degree rather than a binary property. Whereas paradigms such as the VWP have been successful in incorporating visual contexts into the laboratory studies, they still fall short of capturing the continuous, dynamic nature of social interactions. For instance, unlike lab-based VWP experiments, everyday communication does not unfold in discrete trial-by-trial segments; rather, it is continuous (Spivey, 2007; Tanenhaus et al., 1995), and requires us to account for macro-processes at the social-interactional level together with cognitive processes within individuals

(Clark, [1996]; Fusaroli et al., [2014]; Rohlfing et al., [2003]). In an interaction, visual context is not disjoint and resets at the beginning of each trial but it is essentially ever present and evolves as the interaction progresses. This limitation of VWP has recently been acknowledged by its proponents, and, being aware of such limitations, they call for more innovative methods through which real-world contextual effects can be studied in interactive settings (McMurray, [2023]; Ryskin & Spivey, [2023]). Although the methodological gap addressed in this paper is narrower than the general challenge of ecological validity, we aim to bring the methodological rigor of VWP into Human–Robot Interaction (HRI) research and thereby propose a framework for studying and testing predictions about online language processes in a relatively controlled yet interactive HRI setup. The immediate issue is that many established paradigms capture fine-grained, time-locked linguistic processes well, but offer only limited opportunities to examine how those processes change when comprehension, production and action unfold within the same joint task. In response, we leverage recent advancements in HRI in the hope that the field of psycholinguistics stays ready and can move further as the developments follow in upcoming years.

How HRI can contribute to study online linguistic phenomena

A core motivation to observe eye movements during continuously unfolding speech stems from the fact that many meaningful processes related to language occur below the level of conscious awareness (Mishra et al., [2013]). Participants are typically unaware of their gaze patterns as they process spoken language in real-time (Tanenhaus et al., [1995]). For example, in a task designed to measure lexical activation patterns across phonological cohorts, participants who hear a word *cap* may briefly look at a picture of *candle* (a phonological cohort), more often than on unrelated images. Critically, when they were asked to report whether they also looked at *candle* briefly, most participants were unaware of such fixations and reported not looking at the distractors by any means (Tanenhaus et al., [1995]). This kind of automatic processing has been consistently observed not only for competitors at lexical (Allopenna et al., [1998]), but also at the semantic (Yee & Sedivy, [2006]) and syntactic levels (Spivey et al., [2002]), making eye-

tracking a robust tool for capturing cognitive processes unfolding beyond conscious control. This allows one to investigate processes within a setup without disrupting the dynamically updating task goal, as opposed to response times that interfere with participants’ cognitive states and impose restrictions on the dynamic goal by pushing participants to act as quickly as they can. At the same time, in interactive settings gaze is not only informative about ongoing linguistic processes, but can also function as a communicative resource for coordination between partners, making traditional outcome measures such as response times less natural and less meaningful in such settings. In a review article, Degutyte and Astell (2021) summarize studies addressing questions such as whether speakers maintain or avert gaze toward their partner during turn initiation or completion, how attention is divided between partner and task, how gaze is distributed across gestures, and how referential communication unfolds under semi-controlled conditions (Degutyte & Astell, 2021, for a review). Whereas these studies have successfully integrated the novelty of the eye-tracker with phenomena such as turn-taking, with some exceptions, they focus more on communicative aspects than on fine-grained temporal dynamics of linguistic processes. It appears that whenever online language-mediated eye movements are studied, one often has to return either to screen-based paradigms, as in traditional VWP studies, or to introduce an opaque divider between participants in order to preserve a certain degree of interactivity and control (e.g., Brown-Schmidt & Tanenhaus, 2008). In rare cases, where such experiments are implemented in real-world settings, the replication proves difficult—an issue that is also reflected in ongoing challenges of methodological transparency and comparability in eye-movement studies of interaction (Degutyte & Astell, 2021). Taken together, the methodological challenge is that interactive paradigms rarely preserve the temporal precision needed to study online language-mediated eye movements while also retaining an actual interaction partner and an ongoing task.

Rapidly evolving fields such as human–robot interaction in combination with emerging technologies might contribute to implementation of experimental setups that would have been infeasible a few decades ago. One of the key advantages of HRI setups is that they allow the inclusion of an interaction partner and, at the same time, a controlled investigation of the “shared” and “individual” aspects without compromising real-time

processual measures. Since a robot’s behavior can be controlled precisely across trials and study participants, it enables researchers to introduce interaction while retaining a degree of experimental control. Crucially, the motivation for using a robot should be methodological rather than purely ecological, since compared with a human confederate, a robot allows more standardized delivery of instructions, more repeatable trial structure, and stable manipulations. Whereas the idea of employing robots for experiments has been longstanding, and the HRI field has been developing with its own goals, primarily aimed at making robots more responsive, adaptable, and socially effective for humans, it also holds untapped potential for contributing to our understanding of language processing (Bateman et al., 2003). This potential goes far beyond simply observing how people behave in front of robots or how they perceive them; rather, it opens a path toward systematically investigating the temporal dynamics of linguistic processing.

Yet, like many fields that remain relatively inert to interdisciplinary integration, HRI is no exception. For example, the long-term goal of HRI has generally been framed as facilitating more natural interactions between humans and robots, ideally approximating human–human communication (Admoni & Scassellati, 2017; Goodrich & Schultz, 2008). Although such an approach is highly beneficial for the development of more sophisticated robots, the potential of such systems to reveal a fine-grained understanding of language processes remains underutilized. For example, most HRI studies continue to focus on how participants respond to robots under different conditions, aiming to tailor and optimize it based on factors such as anthropomorphism (Murali & Riek, 2023), likability (Brawer & Scheutz, 2018), trust (Charalambous et al., 2023), errors (Salem et al., 2015) and how participants would “like” the robot to behave in turn (Niculescu et al., 2019). These notions dominate the HRI field in abundance, but psycholinguistic notions, particularly online language processing are rarely studied. In the context of eye movements, HRI research tends to examine how robot gaze influences participants’ attention or mutual gaze behaviors (Admoni & Scassellati, 2017; Staudte & Crocker, 2009). In developmental robotics, the emphasis is often on finding out the best way to teach the robot (Vollmer et al., 2016). In contrast, investigating online language processes has received comparatively less attention in HRI, where broader goals

such as natural interaction, trust, and adaptation have often taken priority (Goodrich & Schultz, 2008). As a result, the field appears increasingly shaped by the interests of psychology and computer science, with a focus on a classical stimulus-response (SR) approach (von Salm-Hoogstraeten & Müsseler, 2021), and if this trend continues, the field may increasingly prioritize observable outcomes over the investigation of online language processes (Gardner, 1985). One further challenge in an interactive paradigm with robots is the risk of frequent conversational breakdowns. However, rather than viewing these breakdowns as methodological obstacles, we argue that they can serve as an opportunity to study online language processes (Doğan et al., 2022). In the next section, drawing on work on the origins of such breakdowns in social interaction (Albert & de Ruiter, 2018), we outline the utility of such breakdowns for the realization of interactive setups.

The ubiquitous breakdowns in HRI and how to capitalize on them

Ambiguities are inherent in all technological systems, whether they concern how a system functions at the operational level or arise while interacting with it (Arend et al., 2024). In conversational systems such as social robots, breakdowns in conversation are a major source of ambiguity (e.g., Kontogiorgos et al., 2021). In such situations, it has been shown that learners seek help from their partners (e.g., Shah et al., 2020; Singh et al., 2025). These breakdowns may originate at the systemic level, for instance when the robot fails to process spoken input due to limitations in automatic speech recognition, or at the interactional level, when the human interlocutor fails to comprehend the system’s output and must actively engage to resolve the misunderstanding (Doğan et al., 2022; Singh & Rohlfing, 2024). For our purposes, let us set the systemic ambiguity aside and instead focus on interactional ambiguities, which are central to our setup. Interactional ambiguities are not peculiar to human–robot interaction, but are also central to investigations of human–human communication. In conversational analysis, they are widely studied under the term “repairs” (Schegloff et al., 1977, p. 361). Repairs refer to side sequences within a conversation that are initiated when a source of misunderstanding is identified in a prior turn and is subsequently resolved

through negotiation. One such form is other-initiated repair, in which the listener identifies a problem source in the speaker’s utterance and initiates repair by highlighting the source of trouble, thereby prompting the speaker to readjust or elaborate on a specific aspect of the utterance (Dingemanse & Enfield, 2015). Viewed from another angle, this process is closely related to referential ambiguity, where the referent itself becomes the unit highlighted by the listener as unclear or underspecified in the given context. For example, when a speaker says, “*Can you pass me the slip?*” and the listener responds “*The clip?*”, the listener identifies the ambiguity in the utterance and points to the likely trouble source, presumably referring to the objects available in the shared visual context. If two competing referents are visually present, such as *Slip* and *Clip*, the repair initiation may take the form of an alternative offer, such as “*Clip or Slip?*”. Conversely, if no referential cues are available, the listener may instead initiate a more general request such as “*Huh?*” or “*What?*”, or may simply look puzzled, thereby signaling misunderstanding.

Thus, referential ambiguities can be realized in different linguistic forms, each of which carries the problem source back to the partner in a way that is partly shaped by the shared context. Other-initiated repairs are not exceptions, but foundational to social interaction, and have been found to occur approximately once every 1.4 minutes as engagement builds up (Dingemanse & Enfield, 2015). Gregoromichelaki et al. (2022, p. 163) suggest that repairs can be understood as forms of “divergence” in conversation, that is, moments where effort is required to sustain the flow of interaction, and thus they offer an opportunity to study language without giving up interactivity. In this respect, repairs are not merely failures, but can be treated as a useful resource to investigate online language processes. In social interaction, individuals use repair moments not only to reveal the source of trouble, but also to display their conceptualization of the joint action by directly or indirectly highlighting the referents under consideration (Albert & de Ruiter, 2018).

We propose that, when combined with eye-tracking, repair sequences become a powerful methodological tool for probing ongoing linguistic processes indirectly without interfering with the dynamic goal. Repair sequences make the trouble source linguistically explicit, while eye-tracking provides a continuous and time-sensitive measure

of how attention is redistributed around that source without interrupting the ongoing task. Because this setup remains grounded in interaction, where visual context is continuously available and dynamically changing, it provides a promising setting for studying online language processing by inducing systematic “divergence” within the emerging interaction. For methodological purposes, repairs are especially useful because they create locally identifiable moments of divergence. Unlike broader conversational phenomena, repair initiations are relatively well-defined in form, closely tied to an immediately preceding trouble source, and often accompanied by rapid visual reorientation. This makes them particularly suitable for time-locked analysis in interactive tasks.

An obvious extension of this paradigm is to allow participants to act on the provided visual context, rather than merely observing it passively, as in traditional on-screen tasks (Huettig & Tanenhaus, 2025). This shift makes the task pragmatically situated, with actions and goals at its center. In the following sections, we highlight the relevance of such implementations and illustrate their potential with a few study examples from our own setup.

Relevance of actions in psycholinguistic experiments

There has recently been a debate over the nature of the task in VWP and whether, while listening to instructions, participants should simply watch the stimuli (look-and-listen) or manipulate and act on them (Huettig & Tanenhaus, 2025). Whereas both versions of VWP tasks have their own merits, social interactions in particular require engagement from both participants, where speakers and listeners must possess some degree of autonomy and actively shape tasks together rather than maintain a purely unidirectional flow (Clark & Krych, 2004; De Jaegher et al., 2010). According to De Jaegher et al. (2010), as soon as partners interact, engagement emerges, begins to “take over,” and acquires a momentum of its own. Interaction also entails comprehension and production at a multimodal level, intertwined and unfolding together in the service of shared cognition (Pickering & Garrod, 2004). Visual world paradigms typically employ either comprehension or production alone, thereby missing the opportunity to engage

participants in an interactive task (Tanenhaus et al., 1995). As a result, they rarely examine these two mechanisms simultaneously in a single setting and instead assign rigid roles from the very beginning (speaker or listener) in order to preserve control. This challenge is amplified not only by speaker and listener roles in dynamic turn-taking (De Ruiter et al., 2006; Levinson & Torreira, 2015), but also by the noisy data generated by eye movements in dynamic contexts (Degutyte & Astell, 2021). In VWP studies, the dynamic nature of context is typically curbed by introducing systematic changes either to the visual domain (when vision acts as a context for language) or to the linguistic domain (when language acts as a context for vision), thereby treating these tasks as separate from one another. For example, when the aim is to study predictive processes in language, the linguistic stimuli (sentences) are manipulated in order to observe their effect on a fixed visual context, whereas when the aim is to study how visual context constrains language processing, the visual context itself is manipulated (Anderson et al., 2011). The aim of isolating one of these mechanisms has therefore led to two design types, each suited to a specific question of interest to the experimenter. However, the two designs, whether language is treated as context for vision or vision as context for language, can in one sense be treated alike, since visual context is always present and actual behavior is largely driven by afforded actions (Gibson, 1979). Thus, from a design perspective, actions are methodologically necessary if comprehension and production are to unfold within the same dynamic context. Such a setup should preserve visiolinguistic continuity, conversational bidirectionality, and *actions* grounded in context.

The limitation stemming from the lack of interactivity has also been acknowledged recently in VWP, particularly in relation to passive stimulus observation (McMurray, 2023; Ryskin & Spivey, 2023). It is generally agreed that production inherently entails the sensorimotor system (Pulvermüller et al., 2005), and if production and comprehension are to come together in a joint task, that task must also include action performance. Crucially, these actions must be interpretable in psycholinguistic terms, such that their connection to linguistic processes is clear rather than arbitrary. For such a setup to be informative for psycholinguistics, the task should satisfy at least three conditions: (1) actions must be interpretable in linguistic terms, (2) the interaction

must allow participants to alternate between comprehension and production, and (3) the design must include moments in which visual context can systematically constrain how trouble sources are expressed.

We therefore argue that a continuous and interactive setup should provide participants with an opportunity to act on the visual context and thereby shape the interaction themselves, while at the same time allowing linguistic processes to be probed indirectly without interrupting continuity. Most visual world experiments manipulate the visual context by changing it from trial to trial, but probing the linguistic context requires a method through which the relevant constituents are made prominent via some indirect intervention. Once actions are integrated into the task, the next challenge is how to make the relevant linguistic dimensions visible without interrupting the continuity of the interaction. To probe the linguistic context, some disruption in the system is needed such that “divergence” from the joint goal takes place (Gregoromichelaki et al., 2022). When divergences occur, a reorientation toward emergent joint goals becomes necessary. In this sense, divergences trigger “solution probing” processes within the shared space of action opportunities (affordances). For example, once engagement starts to “take over” and acquires momentum, introducing disruption into the system is more likely to encourage participants to recalibrate their actions through verbal and nonverbal means, because engagements are often robust against external disruptions (De Jaegher et al., 2010).

Investigating these phenomena therefore requires a holistic approach, one that studies interaction “as a whole” by examining behavior at multiple linguistic levels together with the processes unfolding within individuals (Fischer, 2010). When comprehension and production are brought together in an action task, we must also be able to specify the functions that eye movements may serve under dynamic speaker–listener roles on the basis of previous literature. The following subsection introduces the main analytic gaze functions used in the present framework, summarized in Table I and situated within the task structure in Figure 1.

Functions of eye movements during interaction

Drawing on broader findings in the literature, we now define the possible functions that eye movements may reflect in the present framework (Allopenna et al., 1998; Degutyte & Astell, 2021; Tanenhaus et al., 1995). Table 1 summarizes these distinctions at the analytic level, whereas Figure 1 situates them within the unfolding structure of the task. The figure is not intended as a process model in a strong sense, but rather as a schematic overview of how different gaze functions may become relevant as utterances unfold in a shared visual workspace. Subsequently, we examine the plausibility of these functions by drawing on a few examples from our own implementation in an interactive HRI setup. Crucially, these labels should be understood as analytic and operational distinctions rather than as discrete classes of eye movements. They are used here to organize possible functions of gaze within the task and to generate predictions about when gaze may serve referential integration, partner-directed monitoring, or pre-utterance planning.

Two core functions of eye movements are identified during speaking and listening, grounded in past research on language-mediated eye movements. First, vision and language remain in continuous correspondence, dynamically acting as context for each other, a process that has long been central to research on language-mediated eye movements (Anderson et al., 2011; Spivey, 2007). This ongoing coordination is what allows language and vision to operate as a unit throughout interaction. Broader theoretical debates, such as whether language shapes perception or vice versa, can also be located within this continuum (Wolff & Holmes, 2011). Second, beyond these core processes, there are parallel functions serving participatory (monitoring) and task-related aspects (pre-utterance planning). These functions are sensitive to the dynamic situation in which a participant finds themselves at a given moment of interaction, such as speaker or listener. Crucially, unlike traditional unidirectional paradigms such as VWP, we do not assume fixed situations (e.g., listener); instead, we allow the situation to emerge dynamically as the interaction builds up. This dynamicity is also reflected in unfolding task rather than organized around fixed participant roles as depicted in Figure 1. A simple interaction episode helps illustrate how these functions may arise in the present setting. As the robot begins an instruction such as *Place the red ob-*

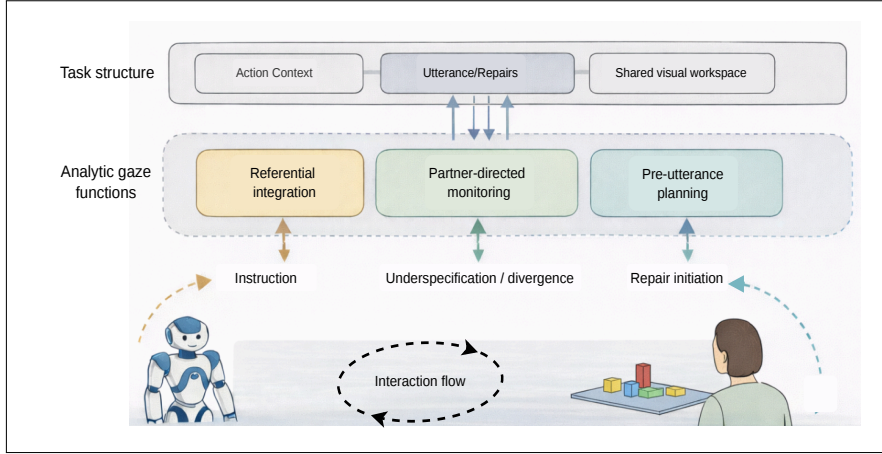


Figure 1.: Overview of gaze functions during interactive task performance. Gaze serves different functions as utterances unfold in a shared visual workspace. Referential integration, partner-directed monitoring, and pre-utterance planning are labeled as analytic distinctions.

ject..., the participant’s gaze may shift from the robot to the relevant object, reflecting *referential integration* of speech with the visual scene. If the instruction then becomes underspecified, for example by omitting the relevant action dimension, gaze may return to the robot as a cue of coordination difficulty (partner-directed monitoring), and before producing a repair, the participant may simulate action-relevant aspects in the workspace to formulate the trouble source more concretely. In this sense, the gaze functions summarized in Table 1 are not treated as isolated, but analytic distinctions that become useful at different points in the interactive sequence.

Studies investigating online processes suggest that when a participant acts as a listener, gaze can serve at least two observable functions. First, it may support the processing of the referent as it becomes visually available, which we refer to as the *referential* function (Altmann & Kamide, 1999; Hanna & Brennan, 2007). Second, it may communicate whether the unfolding instruction is being understood, misunderstood, or requires coordination, which we refer to as the *intentional* function (Clark & Krych, 2004; Kingstone, 2020). In the present framework, the referential function is most closely associated with the listener’s integration of utterance and scene, whereas the intentional function concerns partner-directed monitoring and coordination. A substantial body of evidence shows that listeners typically shift their gaze toward a referent within approximately 200–400 milliseconds after its onset (Allopenna et al., 1998).

Table 1.: Overview of gaze functions discussed in the present framework. The labels summarize analytic distinctions used to organize possible functions of gaze during interaction.

Text-level distinction	Task-level interpretation	Typical role in the interaction	Illustrative timing
Referential	Referential integration of utterance and visual scene	Typically associated with the listener during instruction following, when gaze shifts toward the relevant object as linguistic input unfolds	Listener looks toward the referential target shortly after referent onset
Intentional	Partner-directed monitoring	When gaze functions to signal understanding, misunderstanding or alignment	Listener returns gaze to partner when instruction is underspecified
Simulation	Pre-utterance planning	Typically associated with the speaker before repair initiation or verbalization	Simulation of action-relevant objects in the pre-repair window

On the speaker’s side, eye movements not only exploit visual context for the speaker’s own purposes, but can also contribute to grounding the partner in the shared referential space. Among other modalities, this is often reflected in looks to relevant referents before they are mentioned (Griffin, 2001; A. S. Meyer et al., 1998). In the present framework, such gaze is treated as potentially *partner-directed* when it serves coordination, but as *simulation* when it supports the speaker’s own planning. The distinction matters because production is not only supported by outward signaling, but before a referent or action can be verbalized, some degree of internal conceptualization must also occur.

We therefore treat *simulation* as a function supporting pre-utterance planning and action-related conceptualization (Griffin & Bock, 2000; Pulvermüller et al., 2005). Unlike intentionality, which is directed toward the partner (Clark & Krych, 2004; Hanna & Brennan, 2007), simulation is inward-facing and supports the speaker’s own planning needs. Empirical findings suggest that speakers begin fixating on referents approximately 800–1000 milliseconds before naming them (Flom et al., 2007; Griffin,

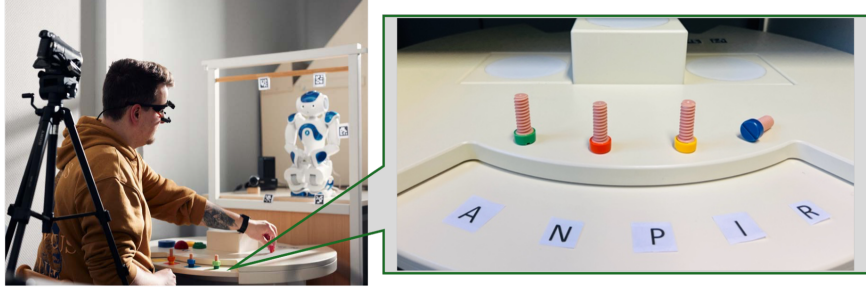


Figure 2.: Experimental setup (left) and example stimuli used in the task (right). Participants interacted with a humanoid robot while wearing an eye-tracker monitoring real-time gaze. The right panel shows the stimuli used in the task, which were to be placed in sequence as instructed by the robot, with a specific manner (vertically or horizontally) and on a designated path (letters).

[2001; A. S. Meyer et al., 1998]. In the present task, this window becomes particularly important because the participant must often formulate a repair in relation to an underspecified action. In such moments, gaze may sample action-relevant aspects of the workspace not only to maintain coordination with the partner, but also to support the speaker’s own conceptualization of the missing action dimension. At the same time, these functions are not cleanly separable in every case. Referential and intentional functions may overlap on the listener’s side, just as intentional and simulation functions may overlap on the speaker’s side. This is one reason why the present article treats them as analytic distinctions rather than as fully discrete categories. Still, the distinctions are useful because they allow us to generate concrete predictions about where participants look, when they look there, and how such looks relate to the unfolding task. In particular, they allow us to link the preceding discussion of action and divergence to the empirical analyses that follow.

Having now outlined how eye movements may serve *referential*, *intentional*, and *simulation* functions within an interactive task, we turn to the question of how these functions can be examined in a concrete HRI implementation. In the following sections, we first introduce the task structure of our setup and then illustrate, through referential and pre-repair analyses, how these gaze functions become visible during comprehension and production in an ongoing interaction.

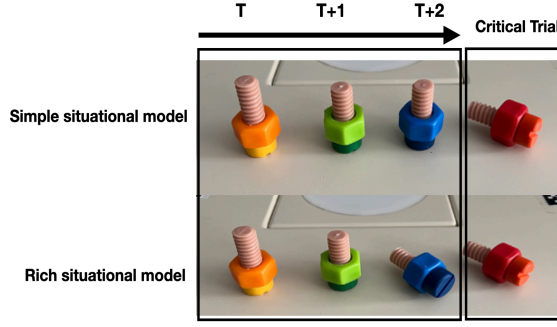


Figure 3.: An example of two situational models made up of a sequence of actions: simple (top) and rich (bottom). Time points labeled T, (T+1), and (T+2) correspond to the action history before the critical trial is presented.

Putting it all together: Studying online language processes in HRI

Figure 2 illustrates our experimental setup adapted from Singh and Rohlfing, 2024. In this setup, a participant wearing a remote eye-tracker (Pupil Labs), interacts with a humanoid robot (NAO) positioned across a table containing physical stimuli. The robot is programmed in a semi-autonomous manner, allowing a flexible interaction by providing resolution to the repairs related to some action dimensions. The critical omissions were predesigned in the robot’s instructions, whereas the participant’s repair initiations and gaze behavior were analyzed as the primary human-dependent measures. Before describing the stimuli and the rationale behind their selection, we first outline the interaction timeline which is central to maintaining interactivity. As discussed earlier, we leverage repair sequences to serve two primary goals in the interaction: a) to bring divergences in the interaction as it builds up, and b) to investigate shared processes by probing indirectly into emerging linguistic behavior for preserving the continuity in ongoing interaction. The full timeline of the interaction flow is illustrated in Figure 4.

Illustrative implementation: task, measures, and analysis

The empirical illustration presented below draws on our recent interactive implementation in which participants followed robot’s instructions while manipulating physical objects in a shared workspace. The task was designed so that the robot’s utterances could establish a referential target and, at critical moments, omit action-relevant in-

formation, thereby creating opportunities for repair. In this way, the setup preserved a continuous joint task while also allowing specific moments of divergence to be introduced in a controlled and repeatable manner. The key design aspect concerned the manipulation of the preceding action context, which allowed us to compare situations in which more versus fewer action possibilities were visually available before repairs were initiated. For the purposes of the present methodological illustration, we focus on two classes of measures. First, we examine the linguistic form of repair initiation, distinguishing between *open* and *restricted* repairs as an indicator of how precisely the trouble source was formulated. Second, we examine eye movements in time-locked windows surrounding instruction onset and repair onset, with particular attention to scanpath distribution across action-relevant areas of interest and to dwell times before repair initiation. Figures 3 and 4 are schematic summaries of the task structure and analytic timeline. Figure 3 shows that rich action context contained all relevant action possibilities before the final critical action, whereas a simple context contained only one action type.

Figures 5 and 6 report empirical patterns from the present implementation. Figure 5 shows aggregate referential gaze dynamics time-locked to instruction onset, whereas Figure 6 shows empirical scanpath and model-based dwell-time patterns in the pre-repair window. Together, these measures allow us to show what kinds of psycholinguistic questions such an HRI paradigm can address and how visual context, repair behavior, and online attention can be related within a single interactive task. The implementation involved 35 participants and 20 critical trials, and gaze was analyzed over predefined areas of interest corresponding to the robot, target objects, and action-relevant locations.

Interaction timeline and design

As the interaction begins, the referential function acts as establishing a shared space, reflected in the integration of language and vision, especially when participants are on listener’s side following instruction. A time-locked window of interest for such referential function is illustrated in the first timeline of Figure 4.

As discussed in the previous section, a “divergence” from the joint goal should take

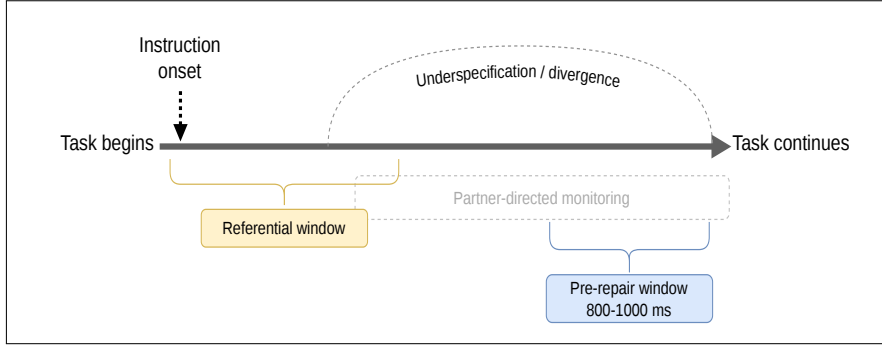


Figure 4.: Timeline of the interaction illustrating key phases. The boxes mark the Time of Interest (TOI), the critical window for analyzing participants’ gaze behaviour during referential, intentional and simulation-related processes.

place to trigger a “solution probing” process. An advantage of such processes is that it does not disrupt the ongoing interaction but becomes a part of it, hence it preserves continuity and provides opportunity to indirectly probe into unfolding linguistic processes at the same time. Critically, when a “solution probing” mechanism concerns linguistic utterance for actions, it calls for simulation (Dingemanse & Enfield, 2015). Since a true “divergence” from interaction can take place only after engagement has already acquired its momentum, we identify this timeline at subsequent stage. After the engagement has taken its form, a critical moment is systematically designed to contain missing or ambiguous information tied to action goals. It brings a disruption in the shared space, where the participants are expected to work together to come to a solution. Since participant and the agent are holding some scale of autonomy for the task, we expect the participant to take a lead for highlighting the source of divergence in mutual understanding from the task goal. The time of interest for these processes should be before the verbalization, specifically within the 800-1000 ms before repair onset, as supported by previous studies (Griffin & Bock, 2000; A. S. Meyer et al., 1998). During repair initiation simulation ensures that the speaker brings into focus the precise source of trouble in the task (in this case, the possible actions) that are needed to be communicated (Dingemanse & Enfield, 2015). It has been proposed that different types of repair encoding have distinct consequences for attentional allocation (De Ruiter et al., 2006; Dingemanse & Enfield, 2015; Schegloff et al., 1977). For instance, as compared to open repair types (e.g., ‘How?’ or ‘Where?’), restricted repair types

(e.g., "Did you mean A or B?") precisely targets the specific trouble source within the ambiguity by demanding more focused attention on the simulated actions (Dingemanse & Enfield, 2015). Although the source of trouble and thereby its verbalization can be uncontrolled, a prudent design where the action elements have a predictable outcome in psycholinguistics can bring some control in the setting. Research in psycholinguistics suggests that any generic action can be decomposed into *path* and *manner* components, a distinction that traces back to the verb-framing typology and the linguistic encoding of motion events (Talmy, 1975, 2000). For instance, in a sentence such as *Put the red object vertically on the letter I*, the term *vertically* expresses the manner of movement, whereas *on the letter I* encodes the path or goal of the action. Prior work also suggests a goal-biasness in event cognition where the path information tends to be more salient than manner, particularly in motion-event cognition (Lakusta & Landau, 2012; Skordos et al., 2020). Hence, capitalizing on linguistically verifiable action elements and how repairs can target them, we structured our experimental design around these ideas to integrate actions. The stimuli thus included four screw shaped objects which could be placed in specific manner (vertically or horizontally) on a designated path as instructed by the robot in a trial. To elicit divergences, the robot occasionally omitted critical information pertaining to either *path* or *manner*. It should be noted that although our implementation was restricted to actions involving vertical and horizontal orientations, the underlying paradigm is by no means limited to these. Rather, the framework can be readily extended to encompass a broader range of actions, provided that their constituent elements can be clearly defined and that linguistic predictions can be formulated to explain their effects. Building upon these foundations, we begin by offering initial insights into how the three functions, *referential*, *intentional*, and *simulation* may be manifested in eye movement patterns when participants act as a speaker and listener during the interactive HRI setting.

Illustrative study: research questions and analytic focus

The empirical illustration below addresses two related questions. First, whether the setup can recover the time-locked referential gaze shifts during robot’s instructions. Second, whether richer action context affects both repair form and pre-repair visual

behavior. To address these questions, we examined repair type and gaze behavior in the time windows surrounding instruction onset and repair initiation. Accordingly, we formulated three predictions regarding gaze functions in a provided action context.

Prediction 1: During robot instructions, gaze should shift from the robot toward the referential target within the expected referential window as depicted in the interaction timeline (Fig. 4).

Prediction 2: Restricted repairs should occur more often in rich than in simple action contexts.

Prediction 3: In the 1000 ms preceding repair onset, rich action contexts should elicit more sustained visual attention on the action-relevant AOIs, especially when participants produce restricted repairs.

Referential

As the interaction unfolds, the robot provides an instruction specifying where and how an object is to be placed. The instructions adhered to a structured template in which the color of the referent was mentioned first, followed by the manner and then the path component (e.g., *Place the red object vertically on letter I*). To examine whether the setup recovers the referential dynamics, we analyzed aggregate gaze time-locked to the onset of the referential expression in the robot’s instruction. Figure 5 shows the resulting referential gaze pattern. The plot summarizes gaze aggregated across trials and participants over predefined areas of interest corresponding to the robot and the objects in the shared workspace. Following referent onset, gaze shifted away from the robot and toward the relevant objects within the expected 200–400 ms window, consistent with classic VWP findings.

This aggregate result shows that the setup can recover online referential processing patterns of the kind traditionally examined in VWP research, while embedding them in an interactive task context. More broadly, time-locked looks in response to the onset of critical words can be extended to multiple linguistic units under investigation, including verbs, nouns, and adverbs. In comparison with traditional offline methods, the advantage of eye-tracking is fully realized when temporal dynamics are analyzed rather than a single point estimate (Spivey, 2007). Crucially, the present setting retains a de-

gree of control comparable to VWP designs while also allowing the continuous timeline of listeners’ referential processes to be visualized as actions unfold in interaction.

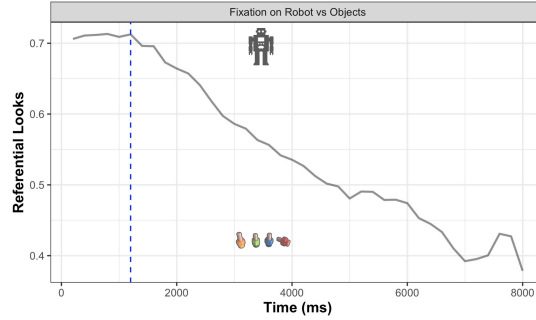


Figure 5.: Aggregate referential gaze dynamics time-locked to instruction onset. The plot shows the empirical distribution of looks to the speaker and task-relevant objects following onset of the referential expression. The vertical blue line marks instruction onset. Gaze shifts toward the referential target emerge within the first 200–400 ms.

Intentional

This section is not presented as a separate empirical test, but as an operational account of how partner-directed gaze cues can be interpreted within the present framework. Although intentionality is a continuous unfolding “macro-phenomena” occurring throughout the task, which serves the core function of bringing joint attention and common grounding within an interaction (Tomasello et al., 2005), it is perhaps one of the most challenging aspects to study using eye movements, particularly in an interactive linguistic task (Holler & Levinson, 2019). Following a functionalist approach, we identified that the intentionality in eye movements is operating all the way from referential processes to maintaining the interaction, running in parallel with linguistic processes (Fig. 4). Whereas listener’s gaze provides cues to the speaker about instruction being (mis)understood, these cues are elicited concurrently as the referent is being processed (Clark & Krych, 2004). Similarly, speaker’s eye movements play a crucial role in establishing common ground for the listener before the relevant referent is uttered (Tomasello et al., 2005). These roles are evident in the speaker’s gaze toward the referent prior to its verbalization (Griffin & Bock, 2000), yet also serving the simulation purposes. Thus, we distinguish intentional functions as serving the partner, i.e., allocentric and referential or simulation functions as serving the self, i.e., egocentric. On one hand, listeners’ gaze toward the referent

while receiving instructions might reflect referential process for conceptualization, on the other hand, a prolonged or perplexed gaze may serve as an indirect intentional cue to the speaker, signaling trouble in comprehension (Singh & Rohlfing, 2024). Similarly, speakers' gaze toward the referent may indicate simulation processes before verbalization. Yet, a sustained fixation on the referent could also function as an indirect cue, implicitly inviting the listener to align their attention with the same referential space in anticipation of the upcoming utterance. In contrast, a direct gaze toward the interlocutor can function as a relatively overt cue relevant to coordination. Thus, intentional gaze cues may serve the partner either overtly (through direct eye contact) or more subtly (through indirect fixation on task-relevant referents). While both forms of intentional cues are likely present throughout the interaction, our current analysis accounts only for direct gaze to the partner as a reliable proxy for intentional cues. In the present framework, looks back to the speaker during repair initiation or task confirmation can be interpreted as direct cues of intentionality communicated through eye movements.

Simulation

An interactive scenario should accommodate both comprehension and production within a unified framework if it is to inform psycholinguistic questions that most unidirectional paradigms cannot easily address. Whereas comprehension research has primarily examined the online integration of speech and visual information, production research has focused more on the encoding of referents and their linguistic realization. Importantly, the conceptualization of an entity to be communicated is shaped by the affordances of the referent in the world (Gibson, 1979). A noun such as *screw*, for instance, may activate object features, but its meaning may also be framed by how it has been acted upon in the unfolding task. In typical VWP studies, where participants do not manipulate objects themselves, such action-related conceptualization may remain implicit and unaccounted. To move beyond this limitation, our setup required participants to act on the objects directly, thereby allowing simulation to unfold in a more grounded manner. Here, we treat simulation as the process by which participants conceptualize actions that are relevant in a given context (Griffin & Bock, 2000; Pulver-

müller et al., 2005). When task-relevant information is omitted, repair initiations can reveal aspects of that simulated action, while also reflecting the participants' commitment to restoring a joint solution. Such processes can be studied systematically only when the relevant action elements are interpretable in psycholinguistic terms rather than being arbitrary. Although action conceptualization itself remains covert, indirect access to its structure can be obtained through repair encoding and the way repairs target the source of trouble (Dingemanse & Enfield, 2015). In this sense, the shared information gap between partners becomes visible during moments of divergence, such that the ensuing "solution probing" process may reflect both the linguistic formulation of the trouble source and the visual context in which it is embedded. If linguistic encoding is indeed shaped by the simulation of actions, then its effects should be observable both in repair format and in eye movements, and crucially, these effects should depend on the visual context available to the participants. To capture processes within this timeline, we focused on the 800–1000 ms window before repair onset, following previous work on pre-utterance gaze in production (Griffin & Bock, 2000) and as illustrated in Figure 4.

To examine this, we compared two visual contexts: *Rich* and *Simple* (Figure 3). A rich action context contained all relevant action possibilities before the final critical action, whereas a simple context contained only one action type. Two dependent measures were analyzed. First, we examined whether repair format differed between simple and rich action contexts. Second, we examined gaze behavior in the 1000 ms preceding repair onset, focusing on scanpath distribution across AOIs and dwell time on action-relevant objects. On this basis, we predicted that simulation would be more prominent when all relevant action possibilities were visually available before the divergence occurred. Critical trials with missing information were embedded within these action sequences, thereby prompting "solution probing" after the interaction had already gained momentum. First, the result suggested that restricted repairs occurred more often in the rich action context, while open repairs occurred more often in the simple context ($\beta = 1.94$, $SE = 0.50$, $p < .001$). This pattern suggests that the available visual context shaped how precisely participants formulated the trouble source. Second, gaze in the pre-repair window differed by context and repair type. In the 1000 ms

before repair onset, scanpaths were more distributed across the visual context when restricted repair was used in the rich situation than in the simple situation (Figure 6 left). The same contrast was reflected in the dwell-time analysis (Figure 6 right), indicating that participants sampled the action-relevant context differently depending on both the available action history and the form of repair they produced (Table 2). More specifically, time-related changes in scanpaths were more distributed across the visual context when restricted repair was produced in the rich situation than in the simple situation, suggesting stronger simulation when the environment afforded a broader set of action possibilities. The pattern was reversed for open repairs: in that case, the rich situation did not yield stronger simulation than the simple situation. This suggests that simulation depended not only on the availability of visual context, but also on how the source of trouble was conceptualized for the partner.

Together, these patterns suggest that structured differences in action history shaped both linguistic choice and visual attention during the pre-repair window. When an informative action model was visually available, participants were more likely to attend to it before producing a restricted repair; conversely, the presence of richer action models was associated with more frequent restricted repair initiations. Methodologically, this finding supports the plausibility of our design for testing how visual context influences language processing, and how language in turn reflects visually available action structure, while preserving the temporality of the interaction. More broadly, the setup shows that HRI can provide a controlled interactive test-bed in which real-time “solution probing” makes it possible to relate action history, repair behavior, and gaze dynamics within a single ongoing task.

Table 2.: Posterior summaries for fixation duration models in simple and rich contexts. Intercepts reflect the reference repair condition. Restr-repair indicates the effect of restricted relative to open repair within each context.

Context	Parameter	Estimate	Est. Error	l-95% CI	u-95% CI
Simple	Intercept	225.44	3.00	219.44	231.34
	Restr-repair	-160.85	6.18	-173.01	-148.57
Rich	Intercept	179.27	8.39	162.60	195.61
	Restr-repair	217.04	16.60	184.07	249.61

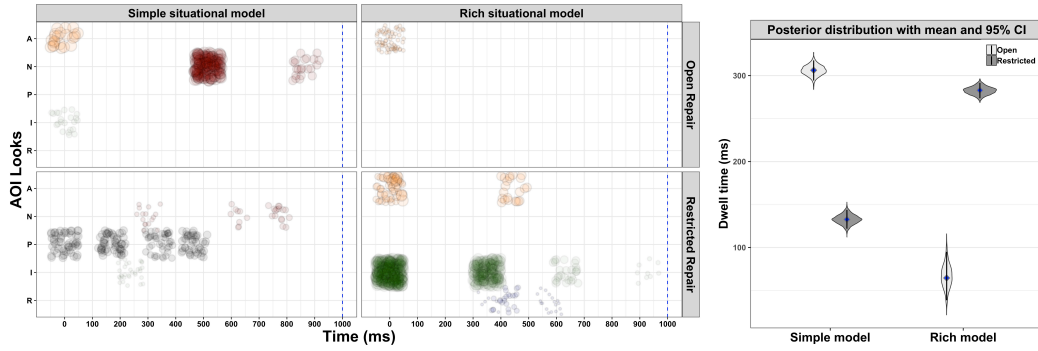


Figure 6.: Empirical visualization of the simulation function of eye movements. The left panel shows gaze scanpaths across situational contexts and repair types in the 1000 ms preceding repair initiation. The right panel shows model-based dwell times for the same conditions, estimated with a Bayesian model (Table 2). The vertical blue line marks repair onset.

Summary and future direction

In this methodological article, we have emphasized the necessity of investigating online language processes through a more interactive lens and have outlined the key factors that have historically constrained the field from fully embracing such approaches. We argue that addressing this gap requires a multidisciplinary perspective that brings together insights from psycholinguistics, human-robot interaction, and cognitive science. Central to the challenge is the integration of two fundamental mechanisms, comprehension and production, within a unified framework that allows for the parallel investigation of their associated online processes. This convergence has long been recognized as a critical condition for achieving genuine interactivity, yet its implementation has often come at the cost of experimental control. Consequently, while a rich body of work has focused extensively on comprehension, particularly through studies of referential processes in language–vision integration, considerably fewer investigations have sought to examine comprehension and production simultaneously within a single, dynamic setup. As a result, to appreciate experimental control, the field has historically favored unidirectional, trial-based paradigms that treat linguistic events as isolated and unrelated, an approach that, while powerful in its precision, falls short of capturing the bidirectional nature of real world language use. We suggested that, whereas one way to move from unidirectionality to bidirectionality is to incorporate production alongside comprehension, another way is to introduce actions into the setting, such that partic-

ipants are not only passively listening or producing utterances but also acting upon objects simultaneously. One effective means of achieving this in a reliable manner is to draw inspiration from real world conversational breakdowns. Repairs, being universal constituents of nearly all conversations, serve the unique function of maintaining the continuity of interaction without disrupting its dynamic flow. We further suggest that integrating all the aforementioned concepts within a human–robot interactive setup can offer a middle ground for investigating online language processes without sacrificing the methodological rigor of psycholinguistics or compromising interactivity. More specifically, the present framework was built around three methodological conditions that we consider necessary if such a setup is to remain informative for psycholinguistics. First, the actions embedded in the task must be interpretable in linguistic terms rather than being arbitrary. Second, the interaction must allow participants to alternate between comprehension and production within the same ongoing task. Third, the design must include moments in which visual context can systematically constrain how trouble sources are expressed. Taken together, these conditions make such a setup both more interactive and methodologically useful for studying online language processes.

While the rationale for employing a humanoid robot to study language may initially seem questionable, it becomes compelling when considered from the perspective of experimental control, allowing for the systematic design of instructions and trials, and thereby offering a powerful tool for introducing targeted manipulations. In light of recent technological developments in AI and robotics, it is an opportune time to integrate these methods into psycholinguistics, particularly to study online language processes rather than merely behavior and personality, as past technological advances have consistently enabled the introduction of innovative methods for examining linguistic processes. The proof-of-concept implementation provided two initial indications that this methodological rationale is workable in practice. First, the setup recovered referential gaze shifts during robot-delivered instructions, showing that time-locked language–vision dynamics of the kind typically studied in VWP can still be observed in an interactive HRI setting. Second, richer action context was associated with more restricted repair initiations and with different patterns of visual sampling in the pre-repair window, suggesting that visually available action structure can shape both lin-

guistic formulation and gaze behavior during ongoing interaction.

We also offer a note of caution when employing such settings: researchers must remain informed by disciplinary findings. Specifically, testable variables should be grounded in robust psycholinguistic evidence, such as the constructs we adopted from motion cognition (e.g., path and manner) or from the psycholinguistics literature on repairs to illustrate our examples. Guided by these theoretical frameworks, we subsequently identified the roles associated with eye movements observed during particular timelines in interaction, namely, *referential*, *intentional*, and *simulation*. We derived these processes from the literature to illustrate how they may manifest within our setting, presenting use-case examples that outline a pathway for investigating specific movements in interaction. The goal here was to avoid unwarranted hypothesis generation and instead ground research firmly in sound conceptual foundations. Crucially, we maintained flexibility in these processes without assigning fixed interactional roles in advance, allowing such roles to emerge dynamically as the interaction unfolded. Ideally, we should be able to discern these roles through the functions of eye movements we have identified, rather than relying solely on the explicit analysis of speech production or comprehension. Nevertheless, the eye movement functions identified as operating in parallel have the potential to be investigated in further studies. An extension of such definitions has already been provided in the context of action learning. Although there is a tight temporal overlap between functions, such as between referential and intentional, and between simulation and intentional, identifying cues that selectively reflect one of these within their complex interplay may offer more specific indicators of action conceptualization and simulation, which in turn may serve as indirect proxies for the engagement of sensorimotor processes. In this sense, action learning mechanisms must be driven by a stronger engagement of the simulation component, and this can be tested through eye movements only when such mechanisms are disentangled from other overlapping functions. We also acknowledge the limitations of our methodology, emphasizing that although the proposed approach has been tested in a specific context by basing our predictions on psycholinguistic literature and eye movement functions observed during interactions, the validity of these functions and the generalizability of the setting must be further evaluated. This includes testing them against other

well-established online linguistic phenomena previously observed, such as real-time semantic integration, referential disambiguation for verbs, and sentence processing. The present article should therefore be read as a methodological proposal with an illustrative empirical implementation. It does not claim that the setup fully resolves the tradeoff between experimental control and natural interaction, nor that the proposed gaze functions can be inferred directly from eye movements alone. Rather, it shows how interactive HRI tasks can be designed so that established psycholinguistic predictions become testable in a more socially situated setting. Nevertheless, we hope this setting offers a valuable test-bed for implementing robust findings from psycholinguistics and for exploring their utility within social interactional contexts.

Data availability

The data and analysis code are available in an anonymized OSF repository (view-only link): [OSF repository](#)

Ethics approval and consent to participate

The study was approved by the Paderborn University ethics committee. All procedures were performed in accordance with relevant guidelines and regulations. Written informed consent was obtained from all individual participants included in the study.

References

- Admoni, H., & Scassellati, B. (2017). Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human-Robot Interaction*, 6(1), 25. <https://doi.org/10.5898/JHRI.6.1.Admoni>
- Albert, S., & de Ruiter, J. P. (2018). Repair: The interface between interaction and cognition. *Topics in Cognitive Science*, 10(2), 279–313. <https://doi.org/10.1111/tops.12339>
- Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for con-

- tinuous mapping models. *Journal of Memory and Language*, 38(4), 419–439. <https://doi.org/10.1006/jmla.1997.2558>
- Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264. [https://doi.org/10.1016/S0010-0277\(99\)00059-1](https://doi.org/10.1016/S0010-0277(99)00059-1)
- Anderson, S. E., Chiu, E., Huette, S., & Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta Psychologica*, 137(2), 181–189. <https://doi.org/10.1016/j.actpsy.2010.06.003>
- Arend, B., Sunnen, P., & Caire, P. (2024). Investigating breakdowns in human–robot interaction: A conversation-analysis guided case study of human–nao communication in a museum environment. *Frontiers in Robotics and AI*, 11, 135978. <https://doi.org/10.3389/frobt.2024.1359782>
- Bateman, J. A., Fischer, K., & Tenbrink, T. (2003). Why a static interpretation is not sufficient in spatial communication. *Proceedings of the 2003 EACL Workshop on Dialogue Systems: interaction, adaptation and styles of management*. <https://aclanthology.org/W03-2702/>
- Brawer, J., & Scheutz, M. (2018). Nonverbal social cues for judging likability in human–robot interaction. *Scientific Reports*, 8(1), 1–9. <https://doi.org/10.1038/s41598-018-25314-x>
- Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive Science*, 32(4), 643–684. <https://doi.org/10.1080/03640210802066816>
- Charalambous, G., Nagy, Z., Fletcher, S. R., & Maciejewski, A. A. (2023). A systematic review of trust assessments in human–robot interaction. *Frontiers in Robotics and AI*, 10, 1172181. <https://doi.org/10.3389/frobt.2023.1172181>
- Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.
- Clark, H. H. (1992). *Arenas of language use*. [Pages: xviii, 419]. University of Chicago Press.
- Clark, H. H. (1996). *Using language*. Cambridge University Press.

- Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1), 62–81. <https://doi.org/10.1016/j.jml.2003.08.004>
- 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](https://doi.org/10.1016/0010-0277(86)90010-7)
- Dale, R., & Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, 56(1), 391–430. <https://doi.org/10.1111/j.1467-9922.2006.00372.x>
- De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Science*, 14(10), 441–447. <https://doi.org/10.1016/j.tics.2010.06.009>
- De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker’s turn: A cognitive cornerstone of conversation. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 7–12.
- Degutyte, Z., & Astell, A. (2021). The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. *Frontiers in Psychology*, 12, 616471. <https://doi.org/10.3389/fpsyg.2021.616471>
- Dingemanse, M., & Enfield, N. J. (2015). Interactive repair and the foundations of language. *Trends in Cognitive Sciences*, 19(11), 654–661. <https://doi.org/10.1016/j.tics.2015.08.005>
- Dingemanse, M., Liesenfeld, A., Rasenber, M., Albert, S., Ameka, F. K., Birhane, A., Bolis, D., Cassell, J., Clift, R., Cuffari, E., De Jaegher, H., Dutilh Novaes, C., Enfield, N. J., Fusaroli, R., Gregoromichelaki, E., Hutchins, E., Konvalinka, I., Milton, D., Rączaszek-Leonardi, J., ... Wilschko, M. (2023). Beyond single-mindedness: A figure-ground reversal for the cognitive sciences. *Cognitive Science*, 47(7), e13230. <https://doi.org/10.1111/cogs.13230>
- Doğan, F. I., Torre, I., & Leite, I. (2022). Asking follow-up clarifications to resolve ambiguities in human-robot conversation. *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 461–469. <https://doi.org/10.1109/HRI53351.2022.9889368>

- Fischer, K. (2010). Beyond the sentence: Constructions, frames and spoken interaction. *Constructions and Frames*, 2(2), 185–207. <https://doi.org/10.1075/cf.2.2.03fis>
- Flom, R., Lee, K., & Muir, D. (Eds.). (2007). *Gaze-following: Its development and significance*. Lawrence Erlbaum Associates Publishers.
- Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, 32, 147–157. <https://doi.org/10.1016/j.newideapsych.2013.03.005>
- Galati, A., & Brennan, S. E. (2021). What is retained about common ground? Distinct effects of linguistic and visual co-presence. *Cognition*, 215, 104809. <https://doi.org/https://doi.org/10.1016/j.cognition.2021.104809>
- Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution*. Basic Books.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.
- Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.
- Goodrich, M. A., & Schultz, A. C. (2008). Human–robot interaction: A survey. *Foundations and Trends® in Human–Computer Interaction*, 1(3), 203–275. <https://doi.org/10.1561/11000000005>
- Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10), 1489–1522. [https://doi.org/10.1016/S0378-2166\(99\)00096-X](https://doi.org/10.1016/S0378-2166(99)00096-X)
- Gregoromichelaki, E., Eshghi, A., Howes, C., Mills, G. J., Kempson, R., Hough, J., Healey, P. G. T., & Purver, M. (2022). Language and cognition as distributed process interaction. *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue*, 160–171.
- Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11(4), 274–279. <https://doi.org/10.1111/1467-9280.00255>
- Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, 82(1), B1–B14. [https://doi.org/10.1016/S0010-0277\(01\)00138-X](https://doi.org/10.1016/S0010-0277(01)00138-X)

- Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57(4), 596–615. <https://doi.org/10.1016/j.jml.2007.01.005>
- Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8), 639–652. <https://doi.org/10.1016/j.tics.2019.05.006>
- Huetting, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, 137(2), 151–171. <https://doi.org/10.1016/j.actpsy.2010.11.003>
- Huetting, F., & Tanenhaus, M. K. (2025). Rethinking task importance in the visual world paradigm. *Brain Research*, 1867, 149965. <https://doi.org/10.1016/j.brainres.2025.149965>
- Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116(1), 26–37. <https://doi.org/10.1037/0096-3445.116.1.26>
- Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.
- Kingstone, A. (2020). Everyday human cognition and behaviour. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, 74(4), 267–274. <https://doi.org/10.1037/cep0000244>
- Knoeferle, P., Crocker, M. W., Pickering, M. J., & Scheepers, C. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, 95, 95–127. <https://doi.org/10.1016/j.cognition.2004.03.002>
- Kontogiorgos, D., Tran, M., Gustafson, J., & Soleymani, M. (2021). A systematic cross-corpus analysis of human reactions to robot conversational failures. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 112–120. <https://doi.org/10.1145/3462244.3479887>
- Kuhlen, A. K., & Brennan, S. E. (2013). Language in dialogue: When confederates might be hazardous to your data. *Psychonomic Bulletin & Review*, 20(1), 54–72. <https://doi.org/10.3758/s13423-012-0341-8>

- Lakusta, L., & Landau, B. (2012). Language and Memory for Motion Events: Origins of the Asymmetry Between Source and Goal Paths. *Cognitive Science*, 36(3), 517–544. <https://doi.org/10.1111/j.1551-6709.2011.01220.x>
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.
- Levelt, W. J. M. (2013). *A history of psycholinguistics: The pre-chomskyan era*. Oxford University Press.
- Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Trends in Cognitive Sciences*, 19(1), 26–33. <https://doi.org/10.1016/j.tics.2014.11.007>
- MacWhinney, B. (2000). *The chldes project: Tools for analyzing talk*. Lawrence Erlbaum Associates.
- Maselli, A., Gordon, J., Eluchans, M., Lancia, G. L., Thiery, T., Moretti, R., Cisek, P., & Pezzulo, G. (2023). Beyond simple laboratory studies: Developing sophisticated models to study rich behavior. *Physics of Life Reviews*. <https://doi.org/10.1016/j.plrev.2023.07.006>
- McMurray, B. (2023). I’m not sure that curve means what you think it means: Toward a [more] realistic understanding of the role of eye-movement generation in the visual world paradigm. *Psychonomic Bulletin & Review*, 30(1), 102–146. <https://doi.org/10.3758/s13423-022-02143-8>
- Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, 66(2), B25–B33. [https://doi.org/10.1016/S0010-0277\(98\)00009-2](https://doi.org/10.1016/S0010-0277(98)00009-2)
- Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2), 227–234. <https://doi.org/10.1037/h0031564>
- Mishra, R. K., Olivers, C. N. L., & Huettig, F. (2013). Chapter 8 - Spoken language and the decision to move the eyes: To what extent are language-mediated eye movements automatic? In V. S. C. Pammi & N. Srinivasan (Eds.), *Decision Making* (pp. 135–149, Vol. 202). Elsevier. <https://doi.org/https://doi.org/10.1016/B978-0-444-62604-2.00008-3>

- Murali, A., & Riek, L. D. (2023). Measuring anthropomorphism in human–robot interaction: A scoping review. *International Journal of Social Robotics*, 15(3), 425–446. <https://doi.org/10.1007/s12369-023-01014-z>
- Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, 5(6), 690–702. <https://doi.org/10.3758/BF03197357>
- Niculescu, A., van Dijk, B., & Nijholt, A. (2019). Users’ preferences for human-like robot speech and gestures: Individual and cultural variability. *JMIR Human Factors*, 6(2), e12629. <https://doi.org/10.2196/12629>
- Piantadosi, S. T. (2014). Zipf’s word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5), 1112–1130. <https://doi.org/10.3758/s13423-014-0585-6>
- 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>
- Pulvermüller, F., Hauk, O., Nikulin, V., & Ilmoniemi, R. (2005). Functional links between motor and language systems. *European Journal of Neuroscience*, 21(3), 793–797. <https://doi.org/10.1111/j.1460-9568.2005.03900.x>
- Reddy, V., Markova, G., & Wallot, S. (2013). Anticipatory adjustments to being picked up in infancy. *PLOS ONE*, 8(8), e65289. <https://doi.org/10.1371/journal.pone.0065289>
- Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination. *Psychological Science*, 18(5), 407–413. <https://doi.org/10.1111/j.1467-9280.2007.01914.x>
- Rohlfing, K., Rehm, M., & Goecke, K. U. (2003). Situatedness: The Interplay between Context(s) and Situation. *Journal of Cognition and Culture*, 3(2), 132–156. <https://doi.org/10.1163/156853703322148516>
- Ryskin, R. A., & Spivey, M. J. (2023). Toward sophisticated models of naturalistic language behavior: Comment on “beyond simple laboratory studies” by a. maselli et al. *Physics of Life Reviews*, 46, 81–83. <https://doi.org/10.1016/j.plrev.2023.09.004>

- Salem, M., Lakatos, G., Amiridze, N., Schroeder, C., & Kopp, S. (2015). The impact of robot errors on trust: A study in home care. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 331–338. <https://doi.org/10.1145/2696454.2696490>
- Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2), 361–382. <https://doi.org/10.2307/413107>
- Shah, J. A., Gluck, K. A., Belpaeme, T., Koedinger, K. R., Rohlfing, K. J., van der Maas, H. L. J., Van Eecke, P., VanLehn, K., Vollmer, A.-L., & Yee-King, M. (2020). Task instruction. In K. A. Gluck & J. E. Laird (Eds.), *Interactive task learning: Humans, robots, and agents acquiring new tasks through natural interactions* (pp. 170–190, Vol. 26). MIT Press.
- Singh, A., & Rohlfing, K. J. (2024). Coupling of task and partner model: Investigating the intra-individual variability in gaze during human–robot explanatory dialogue. *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, 218–224. <https://doi.org/10.1145/3686215.3689202>
- Singh, A., Wrede, B., Rohlfing, B., Groß, A., & Rohlfing, K. J. (2025). Manners matter: Action history guides attention and repair choices during interaction. *Proceedings of the IEEE International Conference on Development and Learning (ICDL)*.
- Skordos, D., Bungler, A., Richards, C., Selimis, S., Trueswell, J., & Papafragou, A. (2020). Motion verbs and memory for motion events. *Cognitive Neuropsychology*, 37(5-6), 254–270. <https://doi.org/10.1080/02643294.2019.1685480>
- Spivey, M. J. (2007). *The continuity of mind*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195170788.001.0001>
- Spivey, M. J. (2023). Cognitive science progresses toward interactive frameworks. *Topics in Cognitive Science*, 15(2), 219–254. <https://doi.org/10.1111/tops.12645>
- Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45(4), 447–481. [https://doi.org/10.1016/S0010-0285\(02\)00503-0](https://doi.org/10.1016/S0010-0285(02)00503-0)

- Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General*, 112(1), 1–36. <https://doi.org/10.1037/0096-3445.112.1.1>
- Staudte, M., & Crocker, M. W. (2009). Visual attention in spoken human-robot interaction. *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 77–84. <https://doi.org/10.1145/1514095.1514110>
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592. <https://doi.org/10.1073/pnas.0903616106>
- Stratou, G., Putze, F., Beetz, M., & Schultz, T. (2021). A humanoid robot increases human language production in a narrative task. *Scientific Reports*, 11, 16617. <https://doi.org/10.1038/s41598-021-95645-9>
- Talmy, L. (1975, December). Semantics and Syntax of Motion. In *Syntax and Semantics volume 4* (pp. 181–238). Brill.
- Talmy, L. (2000). *Toward a Cognitive Semantics* (Vol. 2). Retrieved July 16, 2024, from <https://mitpress.mit.edu/9780262700962/toward-a-cognitive-semantics/>
- Tanenhaus, M. K., & Brown-Schmidt, S. (2008). Language processing in the natural world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493), 1105–1122. <https://doi.org/10.1098/rstb.2007.2162>
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634. <https://doi.org/10.1126/science.7777863>
- Tanenhaus, M. K., & Trueswell, J. C. (2006). Eye movements in spoken language comprehension. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd, pp. 863–900). Elsevier.
- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5), 675–691. <https://doi.org/10.1017/S0140525X05000129>
- Vollmer, A.-L., Wrede, B., Rohlfing, K. J., & Oudeyer, P.-Y. (2016). Pragmatic Frames for Teaching and Learning in Human–Robot Interaction: Review and Challenges. *Frontiers in Neurorobotics*, 10. <https://doi.org/10.3389/fnbot.2016.00010>
- von Salm-Hoogstraeten, S., & Müsseler, J. (2021). Human cognition in interaction with robots: Taking the robot’s perspective into account. *Frontiers in Psychology*, 12, 723356. <https://doi.org/10.3389/fpsyg.2021.723356>
- Wei, Y., Wan, Y., & Tanenhaus, M. K. (2024). Spontaneous perspective-taking in real-time language comprehension: Evidence from eye-movements and grain of coordination. *Scientific Reports*, 14(1), 8031. <https://doi.org/10.1038/s41598-024-58699-z>
- Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. *WIREs Cognitive Science*, 2(3), 253–265. <https://doi.org/10.1002/wcs.104>
- Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(1), 1–14. <https://doi.org/10.1037/0278-7393.32.1.1>
- Yule, G. (1996). *The study of language* (2nd). Cambridge University Press.
- Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, 23(4), 1028–1034. <https://doi.org/10.3758/s13423-015-0864-x>