# Perceptual simulation during linguistic and non-linguistic processing of motion events

## A blank screen eye movement study

DOCTORAL DISSERTATION

submitted to the

Faculty of Modern Languages at Heidelberg University

by

Danny Leander Dirker

First supervisor: Prof. Dr. Christiane von Stutterheim

Second supervisor: PD Dr. Johannes Gerwien

Third evaluator: Dr. Monique Flecken

Defended on December 3rd, 2025, in Heidelberg

# Table of contents

# List of tables and figures

# 1. Introduction

People naturally talk about experiences. It is not uncommon to listen to someone describing, e.g., this new Italian restaurant in town. They emphatically recommend the delicious pizza, recounting how the melted cheese stretched every time they picked up a piece, how they almost burned their tongue because it was oven fresh, marveling at how the Mediterranean herbs mingled perfectly with the sweet-salty marinara, and how the crust was firm but fluffy. Even though no pizza is present now, you imagined the smell and taste of pizza, the satisfaction of eating something delicious, and are likely left with a sudden feeling of hunger, stomach rumbling, or increased salivation — almost as if you were about to eat that pizza yourself.

When we use language, information that is represented linguistically co-activates representations from multiple sensory modalities, like vision, olfaction, audition, or proprioception. This can cause somatosensory experiences of situations that were merely verbally described to us. Scientific research has undertaken dedicated efforts to illuminate this phenomenon, and discussions revolve around the question how we cognitively manage this interplay of sensory modalities with language. Research has long debated whether these co-activations are automatic or may even be necessary to understand or produce language correctly. It is to this debate that the present study contributes.

## 1.1. Problem statement

The controversy is approached from different positions. Cognitive scientific theories from the 4E spectrum (Newen et al., 2018), for instance, take cognition to be Embodied in the sensorimotor system, Embedded in the environment, Enactive through active engagement with the environment, and Extended to external objects. The positions range from an assumption of strong reliance of speaking and thinking on sensorimotor knowledge (Gallese & Lakoff, 2005; Fuchs, 2018; Barsalou, 1999) to more hybrid models that posit concurrent activation of non-linguistic representations alongside more symbolic ones (Barsalou et al., 2008; Louwerse & Jeuniaux, 2008; see Dove, 2009; Dove et al., 2022). At the core of all these views lies the claim that the representations in which our thoughts take shape cannot be detached from experience (Harnad, 1990; Searle, 1992; Barsalou, 2008; Kiefer & Pulvermüller, 2012).

Opposed to 4E approaches, critics counter that cognition based on sensory representations was too computationally demanding to rival the efficiency of symbolic, language-like cognitive processing (Fodor, 1975; Pylyshyn, 2002) and that it is especially ill-equipped to represent abstract thought (Mahon & Caramazza, 2008). It is undisputed that language use inherently depends on specialized cognitive competences. Speakers and listeners apply systematic, combinatorial rules (e.g., syntax) to elements of the mental lexicon in ways not found in other experiential or intellectual domains, causing early cognitivist theories to distinguish language comprehension and production from more general cognitive abilities (cf. Chomsky, 1965; Jackendoff, 1996). This is supported by specific evidence from brain-lesioned patients that demonstrates selective impairment of language use without detriment to other cognitive processes (Dronkers et al., 2004), suggesting that neurocognitive processing of language may in fact be independent of non-linguistic modal systems. Despite this, neuroimaging repeatedly showed activations of sensorimotor brain regions that, though deemed unnecessary for language use, in fact contribute to the resolution of linguistic tasks (Pulvermüller et al., 2005). These competing findings raise the question of when, how, or why sensory-based representations become co-activated with linguistic representations during conceptual processing.

### 1.2. Probing conceptual representation: the present study

The present study takes a step toward resolving this debate. It probes a tenet of grounded cognition that perceptual simulation (Barsalou, 2008) is a core process for conceptual representation, implying that representations of verbal utterance meaning — whether in production or comprehension — are the same as those for the interpretation of the perceptual experience referred to in the utterance. The study is thus guided by the research question whether representations generated by perceptual simulation are systematically co-activated to drive conceptual representation during language comprehension and language production. To investigate this, high-resolution eye-tracking data will be related to underlying cognitive processes of meaning construction, premised on the rationale that eye movements on a blank screen serve as a window into conceptual representation.

For the empirical investigation, 42 participants completed two blank screen eye-tracking experiments in which motion events were presented as auditory stimuli. Experiment 1 examined comprehension of these event stimuli, presented either as non-verbal environmental sounds or as verbal event descriptions, and compared the eye movements between the non-verbal and verbal condition. Experiment 2 explored language production of motion event descriptions, with the environmental sound stimuli serving as prompts. Stimuli of the critical condition referred to motion events in which entities typically change position either in the horizontal or the vertical axis. In both experiments, the aim was to examine whether the comprehension or production activity influenced the directionality of eye movements according to the directionality of the motion events. This design makes it possible to test the above research question whether perceptual simulation drove conceptual representations of motion events, both when they were comprehended as input in verbal versus non-verbal modal format, and when they constituted the message in the language production process.

### 1.3. Main aims and contributions

The present project advances theoretical debates in cognitive science about how concepts are activated and represented, which is relevant for understanding meaning construction. By comparing verbal and non-verbal processing as well as production versus comprehension, the findings illuminate the interplay between linguistic and non-linguistic knowledge representations. They also speak to our cognitive flexibility in handling conceptual representation and concept composition — an issue of growing importance given the global rise of widely available and popular language-based artificial intelligence (AI) applications that process text without the possibility of engaging other sensory modalities, posing significant challenges to 4E theories of cognition.

From a methodological standpoint, this study revitalizes a widely overlooked approach for probing online cognition: analyzing spontaneous, open-eye, non-visual gaze movements. By relating eye movements to processing of non-verbal and verbal auditory stimuli, it extends beyond traditional eye-tracking paradigms that rely on video, image, or text-based material. This crossmodal focus yields novel insights into auditory cognition that go beyond our understanding of purely linguistic or sinusoidal stimuli, rendering this study's tasks more closely aligned with our day-to-day apprehension of

the environment outside the experimental laboratory. Ultimately, this work tests a non-invasive, parsimonious method for determining whether, and when, perceptual simulations may underlie meaning construction.

### 1.4. Definitions of central terms

Some theoretically important but controversial terms (Reilly et al., 2025) must be defined. These terms are *representation*, *concept*, *conceptual representation*, *conceptualization* and *non-visual gaze*.

In this thesis, **representation** is defined as a psychological process of mentation by which a specific aggregate of information is held in mind for current cognitive tasks and where the represented content stands in for something that could be present or absent from the immediate environment. As a tool of thought, representation allows for the interpretation or targeted encoding of sensory and symbolic content alike (von Eckardt, 2012: 30f.). In cognitive neuroscience, the term denotes the coordinated activation of neural assemblies whose recurring patterns are systematically interpreted in terms of mental operations. However, scholars are divided about how physiological activity can directly underlie psychological phenomena, i.e., how so-called *grounding* occurs (for review, see Kiefer & Pulvermüller, 2012). While this thesis is mainly concerned with *mental* (vs. neural) representation, the theoretical foundation (*grounded cognition*; Barsalou, 2008) and the methodological rationale assume an overlap between the physiological and psychological levels. Likewise, in a recent glossary paper, Reilly and 51 scholars from a variety of cognitive scientific disciplines (2025: 263) have proposed to view representation as expressing at both cognitive and neural levels.

A **concept** is a unit of knowledge that aggregates and organizes information in a generalized format. Concepts serve as the building blocks of thought, enabling us to interpret sensations or interoceptions, both online and offline, without relearning from scratch every time we encounter novel input. When activated, they make relevant knowledge available and enable matching this knowledge with current input. Concepts are not necessarily atomic, given that more complex concepts (such as DISHWASHER) may become concepts in their own right while still being composed of more elemental ones. In lexical semantics, concepts correspond to word meanings (cf. *signifié* in

Saussure, 1916; see also Lupyan, 2012). See Reilly and colleagues (2025: 252), Barsalou (2017: 12), or Truman and Kutas (2024: 2) for more detail.

When the mind is engaged in **conceptual representation**, it is running an online process of selectively activating information housed in concepts to guide cognition. Importantly, conceptual representation prioritizes only those features necessary for the task at hand, e.g., for perception, categorization, memory encoding, mental imagery, or fundamental steps in language comprehension and production (see Truman & Kutas, 2024; Kiefer & Pulvermüller, 2012; Reilly et al., 2025). If not otherwise noted, conceptual representation and mental representation are used interchangeably.

Though related to the previous notion, **conceptualization** is a specific term coined in psycholinguistic theories of language production (Levelt, 1989) and therein denotes the processing stage that transforms a speaker's communicative intention into a preverbal message, a form of conceptual representation that can interface with language-specific knowledge bases. During conceptualization, macro-planning and micro-planning operations (e.g., segmentation, selection, structuring, or linearization; cf. von Stutterheim & Nüse, 2003) assemble and order the relevant concepts so that the message can be passed along to the formulator for grammatical and phonological encoding (von Stutterheim, 1999). Note that Barsalou (2008; 2009; 2016) uses the term *situated conceptualization* in a way that renders it largely compatible with conceptual representation, without specifying a relationship with Levelt's coinage.

**Non-visual gaze** refers to eye movements, particularly saccades and fixations, that are not overtly intentional and not guided towards concrete targets in the visual field. Non-visual gaze can be observed when individuals are engaged with internal cognitive processes such as thinking, remembering, or imagining. As such, it may reflect a functional coupling between spontaneous oculomotor behavior and cognitive systems like memory, rather than serving a primarily perceptual role (Micic et al., 2010; Ehrlichman & Barrett, 1983). Importantly, non-visual gaze is not commensurate with the eye movements examined in looking-at-nothing studies (e.g., Altmann, 2004), since those are still guided towards a masked external percept. Nonetheless, the blank screen paradigm often used in those studies can be employed to examine non-visual gaze because the lack of external visual input allows oculomotor control to become more receptive to efferent impulses from neurocognitive activity (e.g., mental imagery).

## 1.5. Structure of the thesis

Chapter 2 reviews the state of research and empirical findings on language and environmental sound processing, event representation, grounded cognition, and motion event cognition. After contrasting speech and environmental sound processing, models of meaning construction and representation (Kintsch, 1988; Radvansky & Zacks, 2014) are described and expanded in a detailed discussion of grounded cognition (Barsalou, 2008), which proposed simulation as a core operation of conceptual representation. The chapter also covers Talmy's (2000a, 2000b) account of motion event representation and justifies, based on relevant empirical evidence (e.g., eye-tracking studies), why motion events are a promising object of investigation for examining perceptual simulation through eye movements. It concludes by situating the research questions within the reviewed literature.

Chapter 3 operationalizes these questions into specific hypotheses and lays out the methodological approach in detail. After making a case for the blank screen paradigm, the chapter elaborates the design of the experiments, the verbal and non-verbal stimulus materials, as well as data collection procedures. The first experiment consisted of an encoding-recall task that triggered processes underlying comprehension of environmental sounds and their corresponding speech stimuli. The second experiment elicited verbal descriptions of the environmental sound stimuli, opening a window on underlying language production processes.

Chapter 4 provides an overview of the analysis procedure, details the calculation of the dependent variables *travel distance* and *saccade rate*, and describes the statistical results for each hypothesis. While main effects of motion event directionality on eye movements were found, some puzzling results emerged and called for further exploration in post-hoc analyses.

Finally, Chapter 5 interprets the findings and embarks on an extensive discussion of the theoretical implications and empirical issues raised, including whether the content of simulations can be controlled, the more general cognitive value of blank screen eye movements, as well as whether perceptual simulations can really be necessary for conceptual processing. Remaining limitations are addressed before the thesis concludes in Chapter 6.

# 2. Literature review

Imagine a scenario like the following and take a guess what could have happened between the situations:

(1) *Johanna is sitting at the dinner table.* […] *To see who it is, she gets up and walks towards the front door.*

What could fill the gap? Most likely, you were thinking of an event such as option 1:

(1.1) *Suddenly, she hears the doorbell ring.*

Alternatively, option 2 could have happened:

(1.2) *Suddenly, her mother shouts from down the hallway: "The doorbell rang!"*

Although it is unquestionable that either option reinstates narrative coherence, they imply fundamental differences in Johanna's perceptual experience. Option 1.1 describes that Johanna perceived non-verbal auditory input, the ringing of a doorbell, as the event motivating her to head towards the door. In contrast, option 1.2 implies that she did not hear such ringing, but instead comprehended auditory input as a verbalization of that audible input. In essence, both auditory inputs refer to the same doorbell-ringing and can equally cause initiation of an appropriate action response[1]. Johanna understands that a person must be waiting at the door but concluded this from inherently different types of perceptual information — spoken language and environmental sound.

Evidently, healthy humans can respond to concurrent inputs in flexible ways and are rarely limited by cognitive abilities to interpret them for guiding action or prediction. We usually manage this with unnoticeable ease and reliable precision, even when faced with input from multiple modalities and streams at once, such as while taking notes (motor, visual, verbal) in a lecture where the speaker (auditory, verbal) points at projected slides (visual, verbal).

[1] This holds irrespective of the implicature in the mother's utterance, that the sound of a ringing doorbell comes with a comparable implicit demand for action.

***Overview of the chapter***

This chapter describes the cognitive abilities that allow us to comprehend auditory input in the verbal (speech) and non-verbal modality (environmental sounds). To do so, this chapter follows an order analogue to how comprehension unfolds in time, from speech perception to high-level processes of conceptual representation.

First, spoken language comprehension (Ch. 2.1) (Menenti et al., 2011; Kuperberg & Jäger, 2016; Hagoort et al., 2004) and comprehension of environmental sounds (Ch. 2.2) (McDermott, 2013; Dick et al., 2016) are discussed in turn and eventually contrasted (Ch. 2.3) regarding characteristics of their information-processing. The subsequent section illustrates theoretical accounts of meaning construction from language or experiential input (Johnson-Laird, 1980; van Dijk & Kintsch, 1983; Zwaan, Langston & Graesser, 1995; Zacks et al., 2007; Radvansky & Zacks, 2014) (Ch. 2.4). A matter of debate is how we ultimately synthesize representations from different modalities into a unified and coherent conceptual representation that manifests our experience of comprehension. Theories of grounded cognition (Barsalou, 1999, 2008; Barsalou et al., 2008) propose that conceptual representation relies on constant interaction between modality-specific and amodal systems (Ch. 2.5). One such proposed mechanisms for conceptual representation, perceptual simulation (Barsalou, 2008), is discussed and applied to previous proposals of meaning construction. A seminal framework of motion event conceptualization (Talmy, 2000a; 2000b) is presented and supported with a review of experimental evidence that concerns the cognitive psychological notions and operations described before (Ch. 2.6). The chapter concludes with an embedding of the research questions (Ch. 2.7).

## 2.1. Spoken language comprehension

When engaged in spoken interaction, our main goal is mutual understanding. To achieve this, we employ a variety of cognitive and motor abilities in parallel. On the one hand, we produce language as audible, structured utterances that transmit messages to the addressee (Levelt, 1989). As addressees, we hear and decode these utterances to infer our interlocutor's intended message. Whether speaking or listening, we monitor not only the linguistic outputs but interpret contextual cues (e.g., visible gestures or reactions, abstract non-verbal cues, antecedent information) while generating predictions about the further course of the conversation on a sentential or discourse level (Kuperberg & Jäger, 2016).

Psycholinguistic research has brought forward detailed empirical investigations and testable theoretical models of language production and language comprehension (Kutas & Hillyard, 1980; Kempen & Hoenkamp, 1987; Levelt, 1989; Allopenna, Magnuson & Tanenhaus, 1998; Levelt, Roelofs & Meyer, 1999; Hickok & Poeppel, 2000; Altmann & Kamide, 1999; Hagoort et al., 2004; Indefrey, 2011; Delogu, Brouwer & Crocker, 2019). Increasingly corroborated with neuroscientific evidence, much research aligns on the consensus that for both processes, despite the reverse direction, we employ identical neural (Menenti et al., 2011; Silbert et al., 2014; cf. Kutas et al., 2007; Hagoort, 2017; Pulvermüller et al., 2006) and cognitive structures (Kuperberg & Jäger, 2016; Pickering & Garrod, 2013). The purpose of this section is to focus on the available structures and procedural stages and to illustrate how we turn audible linguistic input into conceptual information. For this study, the most relevant of these operations is semantic representation (for comprehension) or message generation (for production).

Systematic accounts of language production and comprehension generally agree on a hierarchical architecture and representationalist nature of the underlying processing[2] (Levelt, 1989; Davis & Johnsrude, 2003). The to-be-produced message or the incoming linguistic signal undergo procedural stages, at each of which a (phonological, lexical, or syntactic) mental representation forms the basis for specialized processing and transformation steps (cf. Fodor, 1975: 109f.). In contrast to comprehension, the language processing in production proceeds through these levels

[2] *Language processing* is taken as the multitude of cognitive operations that underlie both the comprehension and production of language in spoken and written form.

of representation in the reverse order. Evidence for the psychological reality of these representational strata comes primarily from language production (see Levelt et al., 1999).[3]

### 2.1.1. Processing stages

Spoken language comprehension begins with auditory sensation. Sounds emitting from human vocal tracts enter our ears as acoustic signals in particular frequency ranges and make hair cells in our inner ear vibrate. This physical stimulation is transduced into electrical impulses that are transmitted to the auditory cortex, from which they are further processed. If the impulses were caused by a language we have mastered, *phonological decoding* begins automatically (Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986). In this first stage, we construct a phonological representation by segmenting the bottom-up auditory stream into discrete strings of phonemes. We apply long-term knowledge about our language-specific phonological inventory, syllable structure constraints, and prosodic features in a top-down fashion. With recognition of the first phonemes and in parallel to ongoing segmentation, potential candidates for words are activated in the mental lexicon. These competitors are retrieved and matched with the concurrent phonological stream and lexemes fitting this phonemic string are subsequently selected. If we are tasked with understanding sentences, our ultimate choice of a lexical unit depends on how it combines with others in the utterance context, satisfying syntactic, semantic, and pragmatic constraints. For instance, in German, the strings /du: hast/ could refer to have$_{2.PS.SG}$ or hate$_{2.PS.SG}$ due to homophony. Disambiguation and selection of the intended word is possible by monitoring preceding and subsequent input and eliminating the incompatible alternatives. Mere linear composition of individual surface forms is hence insufficient to infer the speaker's message.

In parallel to lexical retrieval of the first words, the analysis of syntactic structure, termed *parsing*, is initiated. The lexical items are bound into phrases and sentences. Parsing relies on multiple sources of information to construct such syntactic representations. The access to lemma information in the mental lexicon provides, for

[3] This chapter describes language processing primarily from the perspective of comprehension, although language production is also examined in the present experiments. Framing the discussion in this way facilitates a comparison with the comprehension processes described for environmental sound input, as well as with the models of meaning construction outlined in Chapter 2.3. Explanations of specialized mechanisms in language production are provided when necessary.

instance, inflectional information that governs how words relate to one another[4]. On the encounter of a 3rd person subject noun phrase (e.g., *she*), comprehenders expect a finite verb marked with the English 3rd person inflectional morpheme *–(e)s* (e.g., *say-s*) in the syntactic structure. In addition to grammatical functions of words, their sequential ordering and hierarchical organization into phrases, described as constituent structure, contributes to parsing. Comprehenders abide by grammatical principles of constituent-building (e.g., expecting a lemma of the noun-category after a definite article, as $ART_{DEF}$ NP, e.g., *the boy*), the information-structural status of phrases from discourse context, or prosodic cues when they integrate sequences of lexical units into phrases, and subsequently, in the aggregation of phrases into a cohesive syntactic representation. This illustrates that for syntactic parsing, reciprocal flow of information and interaction between linguistic levels of description is just as pertinent as for phonological decoding and lexical retrieval.

Altmann and Kamide demonstrated in a series of eye-tracking studies (Altmann & Kamide, 1999; Kamide, Altmann & Haywood, 2003; Altmann & Kamide, 2007) that the analysis of syntactic structures is temporally closely linked with the extraction of meaning from an unfolding linguistic stimulus, suggesting again an interaction between representational levels in processing. Altmann (2011) reviewed eye-tracking evidence in temporal measures of saccadic onsets onto objects in visual world stimuli. He concluded that goal-directed gaze shifts are initiated as early as 100 ms from word onset in a concurrent spoken linguistic input. Further seminal work (Altmann & Kamide, 1999) demonstrated how relevant this link is for meaning construction from syntactic structures. Looking at visual world stimuli, participants heard sentences such as *The boy will eat the cake*. An analysis of gaze fixation proportions revealed that participants initiated most looks to the visual object of the *cake* already at offset of the verb phrase (*eat*), so before they even heard the direct object noun phrase. Altmann's findings strongly support that semantic processes, such as predictive inferences based on lexico-syntactic structures, influence comprehension before syntactic parsing is finalized (as opposed to *syntax-first* models, where interaction is expected to happen only in stages following syntactic analysis, cf. Friederici et al., 2004).

The previous descriptions point to non-linearity as a crucial feature of language processing. Even though speech is articulated and perceived linearly and one piece at

[4] Note that not all languages of the world realize grammatical relationships with synthetic means. Also, inflections serve not only syntactic but pragmatic or semantic functions, such as relating propositions to time or other experiential categories. This is not important in the context of this study, however.

a time, phonological decoding and lexical retrieval do not run in strictly sequential fashion — in fact, input is continuously fed into the different processing stages described here. For instance, during retrieval of lemma candidates for the first word, phonological decoding of the second or third word may be ongoing. Components of the continuous input are represented on different levels (phonology, morphology, syntax, semantics) and processing on each level can happen independently from another. In language comprehension research, psycholinguistic consensus exists for such stepwise processing activity, termed *prediction* (Altmann & Mirković, 2009; Kuperberg & Jaeger, 2016). Based on different sources of information during spoken interaction, listeners anticipate not just lexico-syntactic structures of utterances before they are overtly verbalized (by, e.g., co-producing language internally, cf. Pickering & Garrod, 2007) but even the speaker's »message by incrementally updating her hypotheses about this message on the basis of each new piece of information as it becomes available« (Kuperberg & Jaeger, 2016: 39).

As cognitive scientists, psycholinguists wonder how we match such inherently verbal, unified syntactic representations derived from language with non-verbal conceptual categories. The central theoretical construct to facilitate the syntax-semantics-mapping, thematic roles, has been met with disagreement, because »there is simply insufficient evidence to conclude that thematic roles as a class constitute core knowledge« (Rissmann & Majid, 2019: 1864). Assigned to syntactic phrases, thematic roles are taken to be the building blocks of semantic representations in that they clarify universal who-did-what-to-whom relationships between participants of events or actions, such as agents, patients, experiencers (usually noun phrases), all of which are related to one another by a predicate (usually $VP_{FIN}$) (Fillmore (1968) spoke of *deep cases*; Chafe (1970) termed them *semantic structures;* Levelt, 1999: 93). Thematic role structures of utterances are noted as formulaic propositions (e.g., WRITE(STUDENT,THESIS(DOCTORAL))) and to make sense of such propositions, we associate them with world knowledge, be it encyclopedic, contextual, autobiographic, or else.

In other words, once the first constituent has been assigned a thematic role, language comprehenders have reached the stage at which the linguistic representation is transducible into a non-linguistic, conceptual representation. Conceptual representation is generally assumed to be a high-level process that is accessible only indirectly through behavioral measurements, and experimental investigations require

careful modeling and interpretation. Analyses of cortical activity during linguistic tasks have informed theory with precise evidence about the time course (Indefrey, 2011; Friederici et al., 2004; Hagoort et al., 2004) and localization and connectivity (Tyler & Marslen-Wilson, 2008; Turken & Dronkers, 2011), but even sophisticated models refrain from explications of the mental processes involved in conceptual representation of verbal input. Such models are discussed in Chapter 2.4.

### 2.1.2. Summary

In summary, how do we decode what others say to us? We process the input by applying specific combinatorial rules and activating knowledge bases in different stages and construct intermediate, transient representations at the phonological, lexical, syntactic or semantic level. The crucial step from verbal input to meaning is conceived, in the traditional view, as a matching of semantic representations to a knowledge base. It is unclear how these representations are employed in our achievement of reference to entities or events.

## 2.2. Comprehension of environmental sounds

Getting to the bottom of example (1.1), this chapter addresses how the mere perception of the doorbell-sound led Johanna to the same conclusion as her mother's call. Hearing researchers use the term *audition* to describe the physiological and psychological faculties that enabled Johanna to do this. In this subdiscipline of perceptual psychology, such a non-linguistic and non-musical acoustic stimulus is described as an *environmental sound* (Vanderveer, 1979; Ballas & Sliwinski, 1986; van Petten & Rheinfelder, 1995; Dick et al., 2016). Vanderveer (1979) formulated the most detailed definition so far and described environmental sounds as

> »any possible audible acoustic event which is caused by motions in the ordinary human environment. (…) Besides 1) having real events as their sources (…) 2) [they] are usually more complex than laboratory sinusoids, (…) 3) [they] are meaningful, in the sense that they specify events in the environment. (…) 4) The sounds to be considered are not part of a communication system, or communication sounds, they are taken in their literal rather than signal or symbolic interpretation« (Vanderveer, 1979: 16-17 as quoted in Lemaitre et al., 2010: 19-20).[5]

Although they do not always capture our attention, environmental sounds are omnipresent — be it sounds of nature, traffic, people, or household appliances, or else. Johanna's example (1.1) suggests that the ringing of the doorbell was an isolated sound in an otherwise silent environment, however, it is likely that she heard the ringing through a mixture of other sounds. For instance, the low hum of her refrigerator, her above neighbor's footsteps, or the pulsating drone of a distant freight train. Thus, the auditory input, in which potentially meaningful environmental sounds are embedded, reaches our ears as a single stream of sounds that is emitted from multiple sources (Bizley & Cohen, 2013: 694). We must detect and analyze acoustic features in this sensory mix to segregate individual *sound sources*, the »physical entity that generates an acoustic wave« (Alain & Arnott, 2000: D202), if we want to recognize what we hear.

There is myriad neuroscientific research on audition. Influential publications of different groups explain the spatial and temporal blueprint of auditory processing, detailing the hierarchical and functional involvement of peripheral structures, large (sub)cortical structures, or miniscule neural populations across a variety of species (Micheyl et al., 2007; Gutschalk et al., 2005; see Bizley & Cohen, 2013, for a

[5] Note that the use of the term *event* in this definition is not compatible with the way it is used throughout this study (Radvansky & Zacks, 2014). This use refers to a physical occurrence that emitted sound waves.

comprehensive review). While this extensive knowledge has been crucial for clinical applications, such as the development of cochlear implants, the neural substrate presents only one facet of a comprehensive explanation of audition. Regarding the cognitive operations, studies gained laboratory evidence with either speech (Darwin & Carlyon, 1995; McDermott, 2009), music (Iverson, 1995), pure tone stimuli (sine waves oscillating at a discrete frequency, Carlyon et al., 2001; Cusack et al., 2004; Bey & McAdams. 2002; white noise, Kaschak et al., 2006), or combinations thereof (Alain & Arnott, 2000), all of which differ strongly from environmental sounds in ecological validity and informativity, let alone in how they are apprehended cognitively by participants outside of the laboratory (Ballas, 1993; McDermott, 2013: 163; see Ch. 2.3 below). The present study does not investigate the neural basis but the mental representation of meaning derived from environmental sounds and, therefore, calls for an explanation of the cognitive operations underlying environmental sound processing.

### 2.2.1. Levels of cognitive representation

Only few attempts have modeled the mental operations underpinning auditory information processing of natural stimuli (Ballas, 1993; Bregman, 1990; Näätanen & Winkler, 1999; Winkler & Schröger, 2015). Within this niche, the mechanisms described in Bregman's (1990) seminal *Auditory Scene Analysis* have been fundamental for modeling how we perceive the sound input of our surroundings, the *auditory scene*, as cognitively informative.

To get from a sensation to a concept, Bregman's (1990) account conceives of auditory comprehension as going through procedural stages. It is initiated, like spoken language comprehension, with the sensation of undifferentiated auditory sensory input and proceeds through a sequence of operations that group acoustic features into a coherent sensory representation. The main task in auditory comprehension is to detect these sensorily distinct but potentially interpretable gestalts in the auditory stream. Perceptual processing of such coherent sensory patterns and interpreting them results in a mental representation commonly referred to as an *auditory object*. A widely accepted definition describes the auditory object as »the percept of a group of sounds as a coherent whole seeming to emanate from a single source« (Alain & Arnott, 2000: D202). Bizley and Cohen (2013) echo this definition in more technical terms: »auditory objects are the computational result of the auditory system's ability to detect, extract,

segregate, and group the spectrotemporal regularities in the acoustic environment into stable perceptual units« (Bizley & Cohen, 2013: 693), although interdisciplinary agreement far from unanimous (Dick et al., 2016: 1121; see Griffiths & Warren, 2004, for discussion).

Importantly, these definitions imply the view that auditory objects do not exist objectively in the world; rather, it is listeners' processing efforts that isolate them as percepts from the undifferentiated auditory sensory stream. Only when spectrotemporal regularities have been perceived in the bottom-up stream, auditory objects may be detected and associated with potential source situations stored in long-term memory and environmental sound comprehension is completed. In this sense, auditory object perception involves an interactive flow of bottom-up and top-down processing ("primitive and schema-based," Bregman, 1990: 397; Bey & McAdams, 2002; Winkler, Denham & Nelken, 2009). Applied to the context of event perception (see Ch. 2.4.3), the mental representation of an auditory object would be comparable to an *event model* (Radvansky & Zacks, 2014).

In the remainder of this section, the mental representations described in environmental sound processing are compared to the representations proposed in Radvansky and Zacks' (2014) theory of event cognition. To avoid misinterpretation of crucial theoretical concepts, differences are explicitly articulated at relevant points in the text to preempt potential misinterpretation.

Critical discussion of the auditory object as a representational entity is warranted when environmental sounds, as opposed to speech or musical sounds, are the object of investigation. For speech sounds, the voice of a unique speaker would be considered an auditory object (as it distinguishes that speaker from another), whereas in a musical concert, it would be the acoustic output of a unique instrument. Human or instrumental voices are easily parsed into auditory objects because they remain relatively stable in the dimension of spectrotemporal features across all contexts. They will have roughly the same discriminating qualities in, say, 10 seconds vs. 10 days, on the phone vs. face-to-face, in a sound-proof chamber vs. in a sold-out soccer stadium.

In contrast, the perceptual complexity of environmental sounds cannot be straightforwardly bound into a single, homogenous representational unit. Many environmental sound auditory objects are hierarchical compositions of 'smaller' sounds that are clearly discernible in perceptual features (e.g., the different sounds bicycles

make when their gears change), and each sound could be an independent auditory object depending on the context (e.g., a mechanic in a bike shop diagnosing shift-malfunction). In a larger context (e.g., hearing a bike changing gears in busy city traffic), however, the composite environmental sound would be an auditory object of its own, as it is distinguishable among others, yet still composed of more basic sounds. Consequently, it is strongly context-dependent what in the auditory sensory stream listeners bind into meaningful auditory objects. Tackling this problem of granularity, various research groups (Dick et al., 2016; Bizley & Cohen, 2013) proposed multiple levels of description for the mental representation of environmental sounds (cf. Table 2-1 for overview).

| ***sensory processing*** | | ***perceptual and cognitive processing*** | | |
|---|---|---|---|---|
| | | ***auditory event*** | ***auditory object*** | ***auditory scene*** |
| raw auditory sensory input | extracting acoustic regularities (intensity, pitch, timbre) | • smallest relevant unit for meaning construction<br>• single distinct sound occurrences in the environment<br>• component of auditory objects, but could be auditory object of its own (context, expertise)<br>• comparable with *event segment* (Radvansky & Zacks, 2014) | • perceptually constructed, coherent entity that emanates from a single sound source, enabling distinction from other sound sources<br>• either a single auditory event (doorbell) or composed of multiple (using keys to unlock a door)<br>• usually what is labeled when identifying sound source<br>• roughly comparable with *event model* (Radvansky & Zacks, 2014) | • more abstract structure in which auditory objects are embedded<br>• can support identification of auditory objects via schema-based inferences |
| | | individual footsteps; chirp of a bird; a bicycle bell; accelerating engines; | a pedestrian walking in heeled shoes; birdsong; a passing bicycle; driving vehicles; | a busy inner-city intersection |

***Table 2-1***: Hierarchical representational units in processing of environmental sounds. Identification of environmental sounds requires sensory, perceptual, and cognitive subprocesses.

Dick and colleagues (2016: 1022) label the smallest, cognitively relevant component in the construction of meaning representations of environmental sounds an *auditory event*[6] and define it as a singular sound incident with individual acoustic attributes and delimited by clear onset and offset boundaries. *Auditory events* are the 'smaller' sounds of which we compose more complex auditory objects. When conceiving the auditory input of the gear-switching bike as a unique, coherent sound source, for instance, the click of the shift lever would be the first auditory event while the last would be the snapping sound of the chain settling onto the new chainring. The

[6] Winkler, Denham and Nelken (2009: 534) speak of *sound events,* whereas Bizley & Cohen (2013: 693) used the term *acoustic events*, which, by Griffiths & Warren (2004: 887f.) is used for this and a broader sense synonymous to *auditory object*.

coherent sequence of auditory events from the first to the last comprise the auditory object, that the bike has just changed gears. This distinction between auditory event and auditory object is commensurate with the one proposed by Alain and Arnott (2000: D202),

> »While the [auditory object] refers to a perception of a sound source and its behavior over time, the [auditory event] is used when referring to the perceptual dimension of hearing a sound that is occurring at a particular time, in a particular space and having particular attributes (e.g., intensity, duration, timbre). The [auditory, D.D.] event can be part of a larger entity, i.e., the auditory object« (Alain & Arnott, 2000: D202).

It is important not to conceive of the term *event* in *auditory event* as commensurable to theoretical concept of *event* in Radvansky and Zacks' (2014, cf. Ch. 2.4.3) terminology. In the hierarchical system proposed in their theory of event cognition, an *auditory event* would be most comparable to what is termed an *event segment*, a distinguishable unit in perceptual processing that encapsulates a temporally cohesive slice of ongoing activity, within which features remain relatively stable before the next boundary is detected. For the remainder of this study, whenever the term *event* is used, it refers to Radvansky and Zacks' (2014) understanding, and *auditory event* will be used for the currently discussed representational unit in cognitive processing of environmental sounds.

We tend to experience passing bikes outside in city traffic, an environment filled with competing, unsolicited background sounds such as idling or accelerating motorized vehicles, weather, speech, construction noise, to name a few. The auditory object associated with the bike is then embedded, on a coarser level of granularity, in a structured *auditory scene*. The auditory scene largely corresponds to our interpretation of the bottom-up auditory input from which we isolated auditory objects (Bregman, 1990). As such, *auditory scenes* have potential to become schematic sources of knowledge that aid our interpretation of the bottom-up input on a finer granularity level.

Dick and colleagues (2016) consider different subgroups of this higher-level representational structure. Auditory scenes could be *backgrounds* in which qualitatively different, expectable auditory objects may be embedded. For example, when sitting on a park bench on a Saturday, typical auditory scenes contain auditory objects like barking dogs, birdsong, giggling kids, speech, and activating our knowledge about this context helps us interpret the auditory input. In *dynamic auditory scenes*, this expectability warps into script-based prediction (Schank & Abelson, 1977). When

hearing the high-pitched whistles of launching fireworks, we predict that explosion sounds will follow. Lending empirical support to such high-level associations in environmental sound processing, Ballas and Mullins (1991) confirmed that context influences the identification accuracy of environmental sounds. Consequently, schematic knowledge structures derived from familiar auditory scenes can facilitate the recognition of individual auditory objects.

### 2.2.2. Extracting bottom-up features

Before discussing the relevance of various levels of representation for cognitive apprehension of environmental sounds, an outline of decoding processes in audition is due. Essentially, sound pattern detection is initially salience-driven (Kayser et al., 2005; see Itti & Koch, 2000 for the original model on visual perception). Bottom-up analysis of an auditory stimulus may progress substantially and produce a pre-attentive gist (Harding, Cooke & König, 2007; Suied et al., 2014; Isnard et al., 2019, Näätanen & Winkler, 1999[7]) before the stimulus captures our attention (Bregman, 1990: 194; Sussmann, Ritter & Vaughan, 1999; Alain, Arnott & Picton, 2001) and processed into an auditory object.

When exposed to auditory input, automatic analysis of its spectrotemporal qualities begins. To detect and segregate auditory objects from the sensory stream, we employ both simultaneous and sequential grouping strategies (Bregman, 1990). *Simultaneous grouping* integrates synchronously occurring acoustic features and thus mainly helps us decide which sounds originate from a unique source. *Sequential grouping*, on the other hand, integrates causally related but temporally separated sound sequences into a coherent auditory object. This is particularly important when the to-be-determined source emits sound in a contiguous, iterative fashion (e.g., ignition of a car engine) or with minimal temporal delays (e.g., hammering a nail into the wall).

Simultaneous grouping probes the input's consistency in pitch frequency (e.g., low vs. high tone), level of intensity (e.g., soft vs. loud), or timbre (Bizley & Cohen, 2013). Timbre, a less clearly defined feature (see Bregman, 1990: 92f.; Iverson, 1995; Griffiths & Warren, 2004), refers to the dynamic tonal qualities that distinguish auditory inputs with otherwise identical pitch and loudness (e.g., the sound of a trumpet and a

[7] Näätanen & Winkler (1999: 854) describe this as »pre-representational«.

violin, or the two-stroke engine sounds of a lawnmower and a motorcycle). Timbre distinctions are captured by technical terms such as attack time or spectral fluctuation, and lead to judgments of acoustic stimuli sounding bright or dull (Iverson, 1995). In this sense, timbre is a simultaneous feature that depends on temporal patterning. Beyond these acoustic attributes, we are sensitive to temporal or spatial cues. Sounds are likely perceived as emanating from one source if their onset is synchronous (e.g., in roofing, when a metal nail is hammered into a wood beam, distinct sounds emanate from the wood beam and the metal) or if they come from the same location (e.g., the roof).

However, most environmental sounds do not unfold instantaneously, and auditory object perception may require longer exposure. When simultaneous grouping cues do not suffice for determining the auditory object, sequential grouping applies (Bizley & Cohen, 2013). This is the case for environmental sounds caused by iterative actions (Gygi, Kidd & Watson, 2004), such as the noise of the rotor blades of a launching helicopter. Every distinct rotation is audible, but due to the spectral similarity of these rotation sounds in a temporally contiguous series, we group them into an auditory object (Näätanen & Winkler, 1999). Studies of voice segregation by individuals without hearing impairment in cocktail-party situations provide evidence that such sequential grouping may not be purely driven by local bottom-up features, but that segregation of speakers may be schema-driven (e.g., by attending to particular voice qualities; Woods & McDermott, 2018).

### 2.2.3. Top-down operations

Ballas (1993) confirmed that acoustic qualities significantly influence the identifiability of environmental sounds. However, as Näätanen and Winkler (1999: 848) note, an »auditory stimulus cannot be fully described by static features alone«. Successful segregation of various inputs from the sensory stream does not entail recognition of the multi-layered auditory scenes from which these auditory objects resulted. Isolated auditory objects merely provide bottom-up information about the components of a more complex situation, and although certain environmental sounds are meaningful as auditory objects (e.g., a doorbell), in many cases meaning is derived from a composite of equally meaningful *auditory events* on a smaller scale (e.g., unlocking a door). Therefore, to comprehend environmental sounds, hearers typically integrate auditory

objects into a coherent structure (Winkler, Denham & Nelken, 2009: 534; McDermott, Schemitsch & Simoncelli, 2013). This integration process tends to be schema-guided (Bey & McAdams, 2002; Bregman, 1990: 397), i.e., top-down driven, in that hearers use long-term knowledge about situational contexts, which is provided by schemata (Schank & Abelson, 1977; Brewer & Nakamura, 1984), to identify auditory objects.

To illustrate this, consider the example of a falling glass (Dick et al., 2016: 1122). The onset of falling and the falling itself, i.e., the inceptive and imperfective stages, are largely inaudible. It is the shattering sounds following the impact that are the audible information to allow speculation about potential source situations. Thus, integration of an auditory object may only commence once the source situation has reached its resultative stage. In top-down fashion, we apply schematic knowledge about how glasses typically break to understand what happened (cf. »amodal completion« in Bizley & Cohen, 2013: 700). In other words, the knowledge we activate to recognize auditory objects and to identify them may be stored in abstract schemata, such as the *event schemata* proposed in event cognition theory (Radvansky & Zacks, 2014).

In addition to top-down influence in cognitive processing of environmental sounds, the previous example demonstrates the incremental nature of the parsing of simultaneous and sequential bottom-up cues (McDermott, 2013). While the instantaneous smash of crystalline material reveals an impact, the subsequent scattering of the shards sounding from the locus of impact informs us in retrospect that these two auditory events are most likely causally linked and components of a coherent auditory object. In parallel, acoustic analysis of the spreading shards reveals that the falling object was a delicate glass (and not a heavy clay vase) that impacted on a hard floor (and not a carpet).

The importance of schema activation also highlights the psychological validity of hierarchical levels of cognitive representation of auditory stimuli. In decoding environmental sounds, one’s ability to accurately disambiguate and recognize the sound source hinges upon correct identification of its »internally coherent constituents« (Winkler, Denham & Nelken, 2009: 532), since auditory objects are formed out of smaller auditory events, and auditory scenes are representations constituted from auditory objects (Bizley & Cohen, 2013: 704; Ballas, 1993, Experiment 3; Ballas & Howard, 1987: 108).

A further argument relates to codability, the interface of environmental sound and language. Verbal labels, the most frequent result of environmental sound

recognition (McDermott, 2013: 150; Dick et al., 2016: 1123), seem to interface with the environmental sound representation on the auditory object or scene level (Peltonen et al., 2001; Ballas & Howard, 1987: 103; van Petten & Rheinfelder, 1995: 486). Auditory events, the smallest components, are rarely translated into verbal code, likely because the information they provide about the sound source is fragmentary. Nonetheless, previous findings (Lemaitre et al., 2010) and data from the verbalization task in this project (Exp. 2, see Chapter 4.5.4) suggested that when participants were not confident in recognition of the composite auditory object but had an educated guess, they verbalized auditory events of which they were certain, validating their interim representational status. Moreover, Lemaitre and colleagues (2010) report that, in contrast to lay participants, sound experts (e.g., trained or professional musicians, audio technicians) relied more strongly on acoustic than semantic features to group different environmental sounds. Technical proficiency provides experts with explicit labels for fine-grained perceptual distinctions, leading to a »different cognitive organisation of knowledge about the sounds« (Lematire et al., 2010: 25; cf. Lupyan, 2012: 8ff.; Lupyan & Thompson-Schill, 2012).

### 2.2.4. Summary

The processing of environmental sound input unfolds through hierarchical representations, beginning with auditory events, the smallest units of auditory perception, which are rapidly parsed from bottom-up spectrotemporal analysis (e.g., sensing the rhythmic friction sounds of rubber wheels on concrete). Auditory events are grouped into auditory objects, composite units that serve to identify and distinguish sound sources from one another in the environment (e.g., a passing bicycle). At the highest level, auditory objects can be integrated into an auditory scene, a more abstract, global representation (e.g., city traffic) that contextualizes concurrent auditory objects.

While the comprehension of environmental sounds heavily relies on bottom-up processing, the essential step of auditory object interpretation is top-down driven by schemata: Schemata allow us to predict or reconstruct inaudible portions of auditory objects, to decide which competing sensory streams of auditory events may have to be parsed together into a coherent auditory object, and organize multiple auditory objects into a complex auditory scene, thereby enabling accurate categorization and

allowing for predictive inferences that extend beyond the immediately heard signal (Winkler, Denham & Nelken, 2009; Winkler & Schröger, 2015).

## 2.3. Comparison of environmental sound and language comprehension

The preceding chapters explored the cognitive mechanisms and processes underlying the two primary capacities that enable human interaction with the world through audition. Although the auditory information necessary for our cognitive system to identify real-world situations is represented in distinct codes (verbal vs. non-verbal acoustic), it arrives through the same sensory channel: the ears. This warrants a comparison of language comprehension and environmental sound comprehension, which is the focus of this chapter.

### 2.3.1. Physiological substrate

The physiology of the peripheral auditory system in humans is indifferent to types of auditory input. Both language and environmental sound reach our ears as sound waves, therefore the physical signal to be transduced into neural impulses is ecologically identical. To be fair, within the audible spectrum that healthy ears are sensitive to (20-20000 Hz), speech occupies a limited pitch range (approximately 100-300 Hz), whereas environmental sounds scatter across the entire spectrum. Furthermore, the timbral cues produced by the human articulatory system (such as consonant phonemes) fall far short of the timbre variance found in environmental sounds. Although the same neuroanatomical structures transmit the low-level signal (the auditory pathway), neuroscientific findings attest differential recruitment of higher-level cortical structures by speech versus environmental sounds (Visser & Lambon Ralph, 2011; Humphries et al., 2001; Noppeney et al., 2008; van Petten & Rheinfelder, 1995).

This is additionally supported by behavioral evidence that perception of speech sounds in auditory scenes receives a rapid attentional amplification over environmental sounds (e.g., faster response times to a speech stimulus in Agus et al., 2010; see also McDermott, 2013). A recent study found that this attentional amplification is even stronger for the native language as opposed to foreign languages (Liang et al., 2025).

### 2.3.2. Relations between form and meaning

A fundamental difference concerns the semiotic relationships that speech and environmental sounds maintain to their real-world referents. Spoken language is encoded into symbols that make sense for linguistic communities because the associations between what is heard and what is meant are conventionalized (cf. Lupyan & Thompson-Schill, 2012). Apart from onomatopoeia, linguistic surface forms bear no intrinsic resemblance to what they mean, that is, to what they represent conceptually. The form-meaning relationship is arbitrary. The opposite holds for environmental sounds. The mapping between an environmental sound and what this sound means is non-arbitrary and concrete. Environmental sounds represent their sound sources in an iconic fashion, in that they are an integral part of the perceptual experience of the source situation.[8] This fundamental difference in form-meaning relationship has several implications.

The information we use to construct coherent meaning representations of each type of auditory input is available to us in different codes. The informative sensory signals of environmental sounds emit directly from a situation in the environment, while auditory objects of human speech merely help to identify a unique speaker, with reference to situations in the environment becoming possible only indirectly via decoding of linguistic symbols. Thus, when processing spoken language, we must analyze the acoustic signal and extract and parse forms from which we infer the message (cf. Chapter 2.1). Environmental sound comprehension requires no such abstract symbol decoding. Both the surface structure and the meaning have to be parsed from the acoustic features of the auditory input. In other words, form and meaning are extracted from the same dimension, requiring no other specialized decoding capacities for comprehension. Environmental sounds refer to a source situation directly, whereas speech stimuli primarily reveal a speaker and only indirectly refer to a source situation.

Another consequence of this tight form-meaning relationship is that environmental sounds cannot represent abstract concepts as efficiently as lexical items. They are inherently restricted to represent real-world situations that are audible. Even if one imagines some abstract concepts (e.g., CHAOS) to be expressible in sound,

[8] In certain situations, this iconicity holds for the human voice as well. For instance, when interpreting a voice as signaling the presence of a person as opposed to an animal.

the pertinent auditory scenes are likely hard to interpret unequivocally. Linguistic utterances express even the most abstract concepts in maximally economical fashion, sometimes with single elements (e.g., ‘Traffic was chaotic.’).

### 2.3.3. Processing constraints

As a rule-governed, generative system with an unlimited symbolic inventory, language clearly has many advantages over environmental sounds in terms of signal informativity. Despite the structural differences, their comprehension is subject to similar processing constraints. Understanding spoken language and environmental sound input can be equally impeded by frequency of occurrence (Bates et al., 2003; Ballas, 1993, experiment 2) or familiarity (Ballas, 1993: 254) of the present input, with higher frequency and familiarity of input improving identification performance. Moreover, the surface forms of words (phoneme strings) or environmental sounds (frequency spectra) may be homophonic and thus not clearly categorizable on first encounter. Accounting for the fact that a particular auditory event can be caused by multiple sound sources (e.g., activating a light switch vs. a ballpoint pen), Ballas and colleagues (1986), using proportions of distinct interpretations, calculated a measure of causal uncertainty ($H_{CU}$) for each item in their set of environmental sounds. Stimuli with low $H_{CU}$ scores were, on average, named faster (Ballas, 1993).

Although spoken language and environmental sound stimuli can be polysemous, words are largely interpreted as conventionalized labels for category-stereotypical perceptual or conceptual features (Lupyan, 2012: 4). Consequently, and homophonic ambiguity aside, verbal labels guide addressees in the categorization of the input. The lack of such labeling-advantage for environmental sounds[9] impacts straightforward categorization. On another note, the length or duration of stimuli has differential effects when they are speech versus environmental sounds. As hinted at above, even short sentences can describe complex events on a timescale from seconds to decades. Longer speech stimuli contain more words and utterance informativity tends to increase with lexeme count. The picture is not as clear for environmental sounds. Depending on the current sound source, the environmental sound might be harder to identify if it is shorter in duration (e.g., ripping apart paper). On the other hand, longer duration might not yield considerable information gain (e.g.,

[9] Exceptions might be alarms, jingles, or sound icons (e.g., error message) on computers.

a flying helicopter). Accordingly, duration of environmental sounds is not correlated with identifiability nor response confidence (Dick et al., 2016: 1124). In sum, the variable effects of these structural features on comprehension of environmental sound and speech stimuli impose limitations on experimental investigations, as they make naturalistic stimuli difficult to control.

In sum, comprehension of both spoken language and environmental sounds is constrained by the frequency and familiarity of the input, ambiguity arising from homophony or source uncertainty, the availability (or lack) of conventional verbal labels to guide categorization, and the variable informativeness of stimulus duration.

### 2.3.4. Top-down influence

A comparison of the parsing mechanisms makes evident that speech and environmental sound comprehension can be characterized as an interplay of bottom-up and top-down operations (Bizley & Cohen, 2013; Ballas & Howard, 1987). The first steps in understanding spoken language are phonological decoding and lexical retrieval, both of which depend on isolating the smallest meaningful, audible units in the bottom-up perceptual stream. Basically, this is the function of simultaneous and sequential grouping during auditory scene analysis. The difference is that in speech, those units are the syllables that make up words and in environmental sounds, they are auditory events.

Language comprehension proceeds with morphosyntactic parsing, i.e., the integration of lexical units into syntactic phrases based on lemma information. In a general sense, this is analogous to sequential grouping of environmental sound input. The purpose of parsing and sequential grouping is the establishment of dependency relations between transient elements that unfold over time (Ballas & Howard, 1987: 108). Relating one word to another is rule-guided in that we apply implicit structural knowledge of grammar or semantics top-down. Consequently, we can quickly resolve ambiguity of homophonic items. Alternative meanings of ambiguous words can usually be eliminated based on their fit into a superordinate phrase or sentence, in the same way that auditory events can be disambiguated due to their temporal conjunction with other auditory events in the auditory scene (Bizley & Cohen, 2013; Alain & Arnott, 2000; Ballas & Mullins, 1991). Similar top-down influence of superordinate structures on subordinate components becomes apparent when phonemes or entire words are

inaudible, as is often the case in natural settings. Lexeme-level representations in the mental lexicon enable the restoration of phonemes that were masked or missing in the bottom-up input (cf. Warren, 1970). For environmental sounds, a similar pattern inference mechanism has been confirmed in neuroimaging and behavioral studies (»amodal completion« in Bizley & Cohen, 2013: 700; »continuity effect« in McDermott, 2013: 160). Impressions of bottom-up continuity despite inaudibility of crucial components attest top-down influence of schematic representations on immediate perceptual processing of input.

A widely agreed upon manifestation of top-down influence is prediction (see Hutchinson & Barrett, 2019 for review). In language comprehension, top-down support may onset early. Addressees may predict remaining semantic or syntactic constituents after being exposed to the first phrases of a sentence (Kamide, Altmann & Haywood, 2003). When hearing environmental sounds, however, listeners may have to extend their bottom-up sampling beyond the first auditory events before top-down predictive inference of the auditory object is possible (cf. the glass-breaking example in Chapter 2.2.3). Still, even after parsing of individual auditory events, the referent participants (e.g., agents, patients) in the source situation, while usually referred to overtly in the verbal input, may remain unidentifiable. Thus, the way prediction may aid the comprehension of environmental sounds is questionable.

### 2.3.5. Summary

The processing mechanisms underlying the comprehension of environmental sounds and of language exhibit considerable overlap. Both rely on dynamic interactions between bottom-up sensory analysis and top-down inferential processes to construct mental representations organized within multi-level hierarchical systems. Moreover, both domains are constrained by factors such as referential ambiguity, frequency of occurrence, and exemplar variability. Crucially, however, linguistic input conveys information via arbitrary, abstract symbols governed by combinatorial rules, whereas environmental sounds supply concrete, experience-based acoustic cues.

On this basis, probing conceptual representations of motion events (Radvansky & Zacks, 2014; Talmy, 2000b) through a direct comparison of environmental sound versus language comprehension constitutes a theoretically plausible approach for elucidating the extent of perceptual simulations in meaning construction.

## 2.4. Theoretical accounts of meaning construction in language and perception

When the input is language, the breadth of information we can infer from a spoken utterance far exceeds what is explicitly verbalized. The surface structure of the utterance initially guides the construction of an integrated working memory representation of a verbalized situation[10]. These situations, real or imagined, are assumed to become enriched with contextual knowledge via abilities like inference, abstraction, or embodiment (see Tillas, 2014; Zwaan, 2016). Philosophers of mind (Katz & Fodor, 1963; Fodor, 1975, 1983; Johnson-Laird, 1980), cognitive psychologists (van Dijk & Kintsch, 1983; Bower & Morrow, 1990; Zacks et al., 2007), and psycholinguists (Morrow, 1985; Zwaan, Langston & Graesser, 1995; Zwaan & Radvansky, 1998) have not agreed on how humans build such conceptual representations and what it is that they represent.

Therefore, the purpose of this section is to outline different accounts of meaning representation and how they have contributed to our understanding of comprehension. This encompasses *situation models* as proposed by Kintsch (van Dijk & Kintsch, 1983; Kintsch, 1988). Situation models are generally accepted as the central construct driving language comprehension (McNamara & Magliano, 2009) and have since been elaborated to account for the vast neuroscientific evidence on multimodality and parallelism of input processing, as in the *event-indexing model* (Zwaan, Langston & Grasser, 1995; Zwaan & Radvansky, 1998; reviewed in Zwaan, 2016) and cognitive psychological accounts (Zacks et al., 2007; Radvansky & Zacks, 2014). Radvansky and Zacks' (2014) theory of event cognition goes a step further with the claim to account for comprehension of all sorts of input, be it language, visual experience or environmental sounds. The discussion will demonstrate that none of these accounts explain how concepts arise in cognition and proceeds to look for such explanations in theories of embodied cognition (Barsalou, 2008; Barsalou et al., 2008) in the subsequent chapter.

### 2.4.1. Situation models as representational units

Major theoretical and empirical work in the wake of the cognitive turn argued that cognition involved the mental representation of external or internal states-of-affairs

[10] The term *situation* does not refer to the notion introduced in Barwise and Perry (1983).

(Fodor, 1983; Johnson-Laird, 1980). Fundamental abilities like perception and introspection were no longer thought to rely exclusively on direct apprehension of transient external stimulation (Gibson, 1979; Neisser, 1976; Fuchs, 2018), but instead that perceivers built an internal construct of the current input that would be held in working memory for concurrent problem solving.

Johnson-Laird (1980), in seminal discussions of syllogistic reasoning, argued that people created *mental models* as such referential representations. The term *referential* refers to the quality of mental models to be internal constructs that stand for described external circumstances. For instance, people establish mental tokens of the referenced entities and represent them in spatial configurations corresponding to the input to aid inferencing. Importantly, »mental models (…) can be constructed on the basis of either verbal or perceptual information« (Johnson-Laird, 1980: 100). Under this view, language comprehension was not merely conceived of as the transformation of utterances into overtly expressed truth-conditional propositions in an abstract »mental language« (Johnson-Laird, 1980: 104; in reference to Fodor's (1975) amodal *language-of-thought*). In Johnson-Laird's (1980) conception, mental models for language comprehension are analogue: the described or perceived entities are represented by mental tokens that stand for those entities, mirroring the real-world situation described verbally and facilitating mapping across modalities. In verbal communication, the comprehender activates concepts and assembles them into a mental model, attempting to recreate and interpret the situation implied by a speaker's proposition(s).

Contemporaries of Johnson-Laird, van Dijk and Kintsch (1983) similarly wondered how such assembly of concepts into a meaning representation would proceed. Van Dijk and Kintsch (1983) developed an elaborate theory of discourse comprehension based on evidence from reading and memory experiments and mainly attempted to model the cognitive processes underpinning discourse comprehension. They proposed that comprehenders construct layered mental representations which correspond to distinct aspects of information. The *surface form* is a fleeting, verbatim representation of the linguistic utterance in morphosyntactic detail. The assertions expressed in the surface form are summarized into conceptual clusters of micro- and macropropositions which capture sentence meaning in the *propositional text base* (via logical meaning postulates, see van Dijk & Kintsch, 1983: 366ff.; Fodor, 1975: 149). Elaborating the macropropositional structure of the text base beyond what was written

or said, comprehenders draw on previous text information, as well as episodic or schematic knowledge from long-term memory and fuse the propositional text base with this information. The result is the *situation model*, a stable mental representation that binds the components of the input into an aggregate structure that can be effectively encoded into memory and updated or reproduced in further discourse (Zacks & Ferstl, 2016; Kintsch et al., 1990). No longer dependent on the original surface form, the situation model representation captures the situations as they are referred to in discourse.

Yet, how precisely situation models are constructed remains unexplained. Procedurally, van Dijk and Kintsch's (1983) account conceives of situation model construction as going sequentially through separate stages: »We simply could not construct explicitly a situation model without the intervening structure of the propositional text representation« (van Dijk & Kintsch, 1983: 343). They align with Johnson-Laird (1980: 108) on the position that the principal representation of meaning adopts a text-analogue, propositional format (van Dijk & Kintsch, 1983: 344f.), based on which further inferences or modal transformations may occur. This fundamental preference for the proposition was echoed in Kintsch's (1988) later formulation of the *construction-integration model*.

Inspired by artificial intelligence (Minsky, 1974) and connectionism (Rumelhart & McClelland, 1986), Kintsch (1988) details the steps of situation model construction. First, in a *construction phase*, a strict bottom-up analysis of the linguistic input activates lexeme and lemma nodes, as well as syntactic relations, yielding a maximal network of several propositional nodes. These propositional nodes may activate associated semantic nodes through spreading activation. Coherent meaning, however, only emerges after completion of a second processing stage, the *integration phase*. From the network of activated propositions, those that share the highest co-activation strength are included in the resulting mental representation. Kintsch's (1988) account suggests that this integrated *propositional text base*, consisting of verbally activated, interconnected amodal concept nodes, suffices to represent utterance meaning. Relationships between the lexical constituents of a sentence are indexed by overlap of propositional arguments and the corresponding verbal predicates. As the network of propositions includes random inferences from long-term memory, construction of an elaborate *situation model* for comprehension is considered optional (Kintsch, 1988:

180). As such, the *construction-integration model* de-emphasizes top-down contributions to comprehension (e.g., predictive inference).

In previous work, van Dijk and Kintsch (1983) had recognized the importance of top-down influence, attesting situation models the quality to interface with semantic memory in that they »instantiate (…) scripts or frames to become the backbone of a situation model« (van Dijk & Kintsch, 1983: 344). Situation models would thus function like working memory instantiations of schemata, which are stored in long-term memory.

Considering that the verbal experimental stimuli in the present project were single sentence spoken utterances, deeper discussion of meaning construction from connected discourse, narratives, or reading is out of place. Van Dijk and Kintsch's (1983) and Kintsch's (1988) contributions to language comprehension have undoubtedly advanced psycholinguistic and cognitive psychological theorizing about text understanding. Subsequent theories of language production and comprehension (Levelt, 1989) reflected their essential idea that a linguistic utterance can be cognitively operated upon on different levels of mental representation (surface form, propositional text base, and situation model). Despite Kintsch's (1988) claim that a bottom-up extracted propositional network sufficiently represents discourse meaning, it is commonly accepted that comprehenders construct mental models or elaborate situation models that reference both explicit and implicit information in narrative texts, spoken discourse, or real-world circumstances (McNamara & Magliano, 2009), supporting cognitive representation of who is doing what to whom, and why, etc. The precise cognitive mechanisms used in the construction of situation models remain unaccounted for in Johnson-Laird (1980), van Dijk and Kintsch (1983), as well as in Kintsch (1988).

### 2.4.2. Change detected: the event-indexing model

Contending Kintsch's (1988) proposal that comprehenders achieve coherent understanding of situations by activating networks of amodal, symbolic propositions, Zwaan and colleagues (1995) shifted the focus on experientially relevant, dynamic situations as the fundamental building blocks of situation models.

The central assumption in their *event-indexing model* (Zwaan et al., 1995; Zwaan & Radvansky, 1998) is that comprehenders of narratives keep track of changes in described situations with respect to salient universal dimensions (time, space,

causality, intentionality and agents). While parsing the first sentence, comprehenders quickly proceed from constituent analysis to the construal of an initial situation model that is composed of referents as entity tokens in an event structure indicated by the finite verb (Langacker, 1986). This initial situation model (the *current model*) contains indices as to who or what acts, where a situation takes place, the temporal framing of the situation, cause-and-effect relationships between entities, or agent's motivations. When the next event in the next sentence is construed, the current model's established dimensions are monitored for substantial change. If the dimensions retain similar qualities or only undergo minor changes, the situation model representation is updated and modified as these changes are integrated. If dimensions change significantly, as per the introduction of new agents or a new spatio-temporal framework, mapping onto the current model fails and construction of a new situation model initiates[11].

Zwaan and colleagues (1995) report evidence that reading speed decreases as a function of situational discontinuity, with effects of number of dimensions as well as the intrinsic qualitative intensity of change in a dimension (e.g., time in Speer & Zacks, 2005). In an early study, Newtson and colleagues (1977) showed that, segmentation on the basis of dimensional changes is applicable to situations shown in movies (Newtson, Engquist & Bois, 1977). Notably, Zwaan and Radvansky (1998) concluded that not all dimensions equally trigger construction of a new situation model but may be prioritized based on current task goals and relevance. For example, if a narrative is focused on where things happen, location changes are prioritized. Similarly, if a story emphasizes chronology, shifts in time become salient event boundaries (Kurby & Zacks, 2008). Altogether, these studies highlight not only the psychological validity of these situational dimensions for constructing meaning but, at the same time, that event representations are adequate candidates for the cognitive structuring of various forms of input.

Overall, the event-indexing model conceives of situation models as representations that are dynamically constructed online and updated over time, with comprehenders actively engaged in monitoring referent continuity for coherence. In contrast, Kintsch's (1988) model proposed that meaning representations instantiate themselves from activations of nodes, defocusing comprehender-driven contributions to meaning construction. Understanding would occur passively to comprehenders, whereas the

[11] This foreshadowed Zacks et al.'s (2007) concept of *segmentation* at *event boundaries*.

event-indexing model suggests that understanding requires sustained attention to what is happening. Linguistic propositions serve as conventionalized cues about how to incorporate referents and situational dimensions during situation model construction. The view that the structure of the mental representation is analogue to that of real-world, experiential occurrences sympathizes with the contention that situation model construction is an operation that involves all cognitive and perceptual domains — a tenet developed and formulated in Radvansky and Zacks' (2014) theory of event cognition.

### 2.4.3. Event models as domain-independent meaning representations

Experimental evidence testing the event-indexing model indicates that comprehenders of language achieve coherent understanding of narratives by building situation models with event representations. Zacks and colleagues (2007; Radvansky & Zacks, 2014) project this idea beyond language comprehension and conceive of event representations as the fundamental building blocks of our thinking about our experiences of the world. In specific, the theory proposes that mental models of events do not just serve as *situation models* in language comprehension but equally as *experience models* in perception. Radvansky and Zacks (2014: 17) unify these theoretically distinct types of mental representation into an overarching *event model* and claim that comprehension of all input, be it non-verbal or verbal, relies on this type of representation. Event models guide our cognitive system through what is happening in our perceptual stream.

In Radvansky and Zacks' (2014) theory of event cognition, an event is defined as »a segment of time at a given location that is conceived by an observer to have a beginning and an end« (Zacks & Tversky, 2001: 17). This broad definition entails, in principle, that events begin to exist once an experiencing subject perceives and conceives it; that an event is delineated by temporal boundaries and inherently contains temporal structure ranging from a start to an end; and that it is restricted to some (internal or external) space. Prior to elaborating the principles and mechanisms of the theory in more detail, a rough sketch of event comprehension sensu Radvansky and Zacks (2014) is in place.

#### 2.4.3.1. Event segmentation

Imagine that your attention is captured by the sound of somebody walking stairs in your house. With the onset of sensory perception, you automatically begin *segmenting* the perceptual stream by detecting changes in the ongoing activity at different temporal resolutions. This way, you successively create perceptual representations of discrete segments of the sensory input that you are experiencing. In parallel, you activate potential schemata from long-term memory to categorize the perceptual input. With continuous input processing ongoing, you check whether the schemata you are activating match your perceptual representations and vice versa. You have found a match when the further input streaming aligns with what you would expect to happen based on the *event schema* you activated. At this point, you create a *working model* as the interface conceptual event representation that stands for what you are currently experiencing. This is the moment when you have comprehended an event. The now active working model allows you to make schema-driven *predictions* as to how the event is going to play out in the imminent future. Your predictions are true for as long as you experience the same event happening. Your predictions are false when the perceived event changes to such a degree that the working model does not predict your perceptual experience anymore (e.g., the person has reached the top of the stairs). When this happens, you notice an event boundary and are experiencing something new. You shift the current working model as a particular *event model* into memory and this process reinitiates.

This sketch outlined the recurrent interplay that underlies event comprehension and touched upon its critical structures and mechanisms. Importantly, these are *segmentation*, the *working model*, *event schemata*, and *prediction error monitoring*.

Event segmentation refers to an automated neural mechanism by which workable-sized, meaningful units of information are extracted from the perceptual stream. The bottom-up input is monitored for continuous activity. If patterns of this activity contrast with previous patterns, change is detected, and this implies the end of a particular event segment, the beginning of another segment, and establishment of a boundary between them (Zacks et al., 2007: 277). Crucially, this does not imply that a new meaningful event began altogether because event segmentation may occur at different levels of granularity. For instance, if somebody climbs stairs, the event may begin with the first step on the stairs and end when both feet touch the ground on the

next floor. Although a stair-climbing event is made up of multiple successive steps, each of which constitute relevant individual segments, it is all segments together that compose the meaningful event of stair-climbing.

A consequence of this hierarchical organization of event representations is that comprehenders can recognize events from smaller integral components and, importantly, before they have ended. As soon as comprehenders have heard a few steps on the stairs, they may undoubtedly recognize them as segments nested within a larger event (Zacks et al., 2007: 273).[12] This is because event representations serve as clustering structures for information in long-term and working memory. To describe the working memory representation, Radvansky and Zacks (2014: 31) chose the term *working model*. Basically, the working model captures the unique situation that is currently unfolding. It is characterized as a transient, activation-based representation of a particular event experienced by the observer in a particular episode. Its purpose is to provide the observer with a conceptual grasp of their experience and support immediate perception, attention and memory encoding with a structured internal representation »that is maintained by recurrent patterns of neural activity« (Radvansky & Zacks, 2014: 31) at the cortical level. Importantly, working models allow observers to make predictions as to how the current event is likely to continue and, therefore, what sensory input can be expected. These predictions are possible because observers store schematic event representations in long-term memory and activate those for top-down facilitation of event segmentation and categorization. These *event schemata* (Radvansky & Zacks, 2014: 51f.), like scripts (Schank & Abelson, 1977[13]) and schemas (Bartlett, 1932/1995; Brewer & Nakamura, 1984[14]), are robust, systematic knowledge structures that contain information about typical sequences within events, recurrent participants, or cause-and-effect relationships, to name a few. Event schemata develop through learning of commonalities across repeated events or classes of events. In neurophysiological terms, they are considered weight-based memory as their neural manifestation comes from lasting changes to synaptic weights (Zacks et al., 2007: 275). During comprehension, observers quickly activate event

[12] In the terminology of environmental sound perception, the stair-climbing would be the *auditory object*, while each step on the stairs would be an *auditory event*.
[13] The »predetermined, stereotyped sequence of actions that defines a well-known situation« (Schank & Abelson, 1977: 41). See also Sowa (1984: 27), where »(a) schema is a pattern for assembling units called percepts«.
[14] The »unconscious mental structures and processes that underlie the molar aspects of human knowledge and skill« (Brewer & Nakamura, 1984: 42)

schemata to make sense of the current perceptual input by using them as general instructions to assemble the perceived event segments into a meaningful working model (Radvansky & Zacks, 2014: 148). More importantly, they allow comprehenders to predict, in top-down fashion, how the current event continues, i.e., to anticipate regular associations of perceptual features, or to become sensitive toward input cues that signal event boundaries. When a prediction fails, attention may be captured or redirected by the perceptual mismatch, helping to detect event boundaries. If the current event finishes, it is moved to episodic memory and remains available offline for reactivation or integration into a larger, overarching event (e.g., the way that chopping onions is an integral segment of a Bolognese-making event). This representation has been termed *event model* (Radvansky & Zacks, 2014: 7). Event models retain traces of unique associations of features of a particular working model (e.g., that the chopped onion was not fresh anymore and may have caused that sour flavor in the Bolognese) and can be reactivated to feed current working models. Since they capture the meaning of a particular episodic state-of-affairs, event models are the conceptual pendant to Zwaan's and Kintsch's situation models (Radvansky & Zacks, 2014: 17): they capture our understanding of a situation. Accordingly, event models can be encoded into long-term memory as an instance or modify already existing event schemata over time through learning.

#### 2.4.3.2. Event models are multimodal

Radvansky and Zacks' (2014) proposal goes beyond Zwaan's (Zwaan & Radvansky, 1998) in that it explicitly takes events as the structuring mental units across all domains, in all memory systems, and in all modalities, making event representations the fundamental building blocks of our experiencing of, attending to, and thinking about the world. This raises the question of what it is that event models contain and which format they assume; in other words, what is the specific content that people think about when they think about events?

Event models necessarily contain information about space and time of the event (Radvansky & Zacks, 2014: 18ff.). In the established spatio-temporal framework, referent entities (agents, objects, ideas) with physical (e.g., size, shape) or internal characteristics (e.g., emotions, goals) are linked to one another along different dimensions (Radvansky & Zacks, 2014: 22ff.). Comprehenders of events represent, for instance, the identity of an object as it moves from one location to another in a

certain period of time and thus make sense of strategic passing maneuvers in a soccer game. This aspect of situatedness not only echoes the dimensions of the event-indexing model but reveals principal characteristics of the format of event representations.

As assemblies of information from different sensory domains and language, event models are multimodal and embodied (Zacks et al., 2007: 274; Radvansky & Zacks, 2014: 50). Multimodality is a prerequisite for event segmentation and perceptual prediction, since both mechanisms require sensitivity to idiosyncrasies in the bottom-up sensory stream and, in parallel, top-down mapping of expectable changes in this perceptual input. Related to multimodality is the quality of event representations to be analogue (cf. Johnson-Laird, 1980), that is, isomorphic to structural features of the input. This does not imply that event models are, say, detailed visual representations of external input, but that they »contain information relevant to understanding the basic structure of the event« (Radvansky & Zacks, 2014: 26), such as spatial or temporal order.

Another characteristic refers to the construction of event models. Given that sensory information from real-world situations or verbal utterances about them unfolds over time and is not available instantaneously, event models are built from consecutive segments in piecemeal fashion and, in contrast to Kintsch's (1988) assumption of propositional text base, meaning does not become available as an all-encompassing representation at once. On top of that, event models mesh with event schemata, relatively holistic knowledge stores, and profile those aspects of event schemata that are necessary or sufficient for comprehending the current input.

#### 2.4.3.3. The role of prediction

Although hinted at previously, the procedural operations that drive the perception of events and the construction of event models require elaboration. The central operation underlying event model construction is *prediction error monitoring*. During perception, the cognitive system automatically tries to predict what happens next within and after the currently unfolding event. A prediction mechanism enables this by matching perceptual input with event-schematic information from long-term memory (Radvansky & Zacks, 2014: 50). If the event does not continue in an expectable way, such as when one activity ends and another begins, our predictions fail, and the system signals an increase in prediction error. When prediction error exceeds a certain threshold, an

event boundary is detected, and our conceptual system is triggered to update the working model. This initiates a gating that opens working memory to bottom-up sensory influx and top-down from concomitant activation of event schemata or previous episodic event models to find a match to the incoming perceptual input (Zacks et al., 2007: 275) – a new event model is constructed. Throughout these processes, prediction error monitoring continues. As long as prediction error keeps returning too large, model construction proceeds. Once prediction error has decreased below a certain threshold, the cognitive system is signaled to settle for and maintain the newly integrated working model. Cortical activity during event segmentation tasks signals a joint effort of several distinct neural structures in support of the prediction and segmentation mechanism (see Zacks et al., 2007: 283-288 for review of this evidence).

At a more fine-grained level, prediction error monitoring causes event segmentation. Before an event model is constructed in response to an activity, the activity is segmented into smaller activities. For instance, the chopping of an onion is made up of multiple iterations of pushing and slicing with a knife through the vegetable. On a small scale, between each of these segments, boundaries are set (Radvansky & Zacks, 2014: 80f.). We know that the chopping-event continues if, after each segment, inherently similar and predictable segments follow. Prediction and segmentation thus turn raw sensory input into workable percepts for cognitive processing.

Despite the plausibility of prediction-driven working model updating for effective comprehension of dynamic activity, the theory conceptualizes this crucial mechanism as based on a rather static representation: »event models are isolated from their inputs, storing essentially a snapshot of the current event« (Radvansky & Zacks, 2014: 51). When an event boundary is detected, the cognitive system establishes a new event model in a rapid, converging flow of information from different sources. Once this new event model is established, a gating mechanism limits the influx of perceptual input, allowing in only as much as is required for effective prediction error monitoring. This enables stable maintenance of the model across the ensuing dynamic activity until its predictions fail at the next event boundary[15] (Radvansky & Zacks, 2014: 34). The advantage of shielding the working model from continuous, potentially interfering input is that a stable representation of the current real-world situation is maintained even when input is temporarily missing. This is exemplified by the phonemic restoration

[15] In earlier work, Zacks et al. (2007: 275) acknowledged that »the influence of event schemata on event models is continuous and unaffected by the gating mechanism. However, this claim is based largely on parsimony and may need to be revised in the future.«

effect (Warren, 1970) in speech comprehension. That is to say, static representations update models of dynamic activity through sequential manifestation of multiple stable models for segments on a smaller time scale.

#### 2.4.3.4. The importance of top-down influence: schema activation

As noted above, the theory of event cognition acknowledges top-down influence of long-term memory structures on working model construction. Event schemata pervade practically all cognitive operations, be it through a supply of predictions or maintenance of a stable working model (e.g., by providing global event structure), by introducing an estimated error threshold (i.e., allowing one to determine when perceptual input mismatches the event schema), through segmentation guidance (i.e., by imposing typical internal structure, a blueprint of the activity over time), or facilitated memory encoding (i.e., storing the current event as a token-realization of the event schema) (Radvansky & Zacks, 2014: 148). In other words, event schemata are quickly activated to contribute the general information to which comprehenders match the currently processed bottom-up input (Radvansky & Zacks, 2014: 51). The automatized nature of this matching-process suggests that event schemata represent structural features of events and avail of, e.g., associated physical information about entities, their movement paths, manner or speed of movement (Zacks et al., 2007: 275; Lindsay, Scheepers & Kamide, 2013; Kamide et al., 2016; Huette et al., 2012). In fact, event schemata »expand the effective capacity of event models by storing predictive information about the future relevance of certain events« (Zacks et al., 2007: 275). Consequently, they alleviate working memory load and support active representation of and attention allocation to relevant event aspects.[16]

#### 2.4.3.5. Prediction is mental simulation

Given the crucial role of event schemata for prediction, some critical questions emerge: How exactly are perceptual predictions derived from event schemata, or simply put, how does one know what perceptual input to expect? How is abstract, schematic

[16] On a side note, the notion of event schema explains the introductory example of Johanna responding to the doorbell. The reason that Johanna knew that somebody was at the door from both the environmental sound and the verbal input equally is that both inputs activated the same event schema. This event schema entails knowledge about appropriate behavioral reactions to doorbells and where in your home you would expect the bell-ringing person to be. This sufficiently explains why Johanna knew that, first, somebody was waiting for her reaction and, second, that this person was located at the front door.

information matched with concrete perceptual representations during prediction error monitoring? What enables targeted selection of an event schema as a match for the current perceptual input? These questions remain unanswered in Radvansky and Zacks' (2014; Zacks et al., 2007) work and foreshadow the research questions of the present thesis.

Despite remaining agnostic about the processes underlying prediction, event cognition sympathizes with the idea that event models are grounded in sensory-motor modalities (Barsalou, 2008, 2009, 2016): »In our view, perception and prediction are tightly interleaved with motor simulation« (Zacks et al., 2007: 288). Aligning with 4E theoretical accounts of 4E cognition, they adopt mental simulation as a computational vehicle for the conceptual representation of events, yet only for cases where direct sensory input of an event is absent – as when parsing events from language (Radvansky & Zacks, 2014: 11, 69f.) or, possibly, when recalling past events. The main idea from 4E cognition that resonates here is the existence of a common, structured representation into which information from all modalities can be integrated, namely event models.

The theoretical proximity of Radvansky and Zacks' (2014) theory of event cognition to theses of 4E frameworks suggests an answer to the above question(s) to be the following: Abstract event schemata are transformed via perceptual simulations to match the perceptual representations built from bottom-up sensory input. Through simulation of relevant top-down information, the observer generates the to-be-expected perceptual features as predictions against which the current sensory input is compared (Barsalou, 2009: 1284). Prediction error occurs if the resulting representation of schema-driven simulation is incommensurable with what is perceived in the sensory stream. Another possibility is that event schemata already encode information in modality-specific formats matched to sensory experience (e.g., as mental images; Kosslyn, 1980), rendering an online simulation mechanism obsolete.

Altogether, Radvansky and Zacks' (2014) theory of event cognition is more welcoming of mental simulation than shunning it. Explications of how simulation may be involved in event cognition are absent in the literature (Radvansky & Zacks, 2014; Zacks et al., 2007). In fact, the position that »the cognitive representations and processes that are used to process events as they are experienced in the world also are used to simulate them mentally« (Radvansky & Zacks, 2014: 11) renders simulation an epiphenomenon and not an integral process of event model construction.

Previously, however, they considered it a by-product of prediction: »processing is oriented in time such that it results in predictions about the future state of perceptual representations. For example, extracting a motion contour leads to predictions about the future locations of objects« (Zacks et al., 2007: 274). This reads as if simulations of immediate perceptual input contributed causally to active understanding. An in-depth discussion of mental simulation follows in Chapter 2.5.

#### 2.4.3.6. Event models in language comprehension

The previous subchapters discussed comprehension models for verbal input (van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998; cf. Chapters 2.4.1 & 2.4.2). Radvansky and Zacks' (2014) theory of event cognition goes beyond those approaches, at least, with the attempt to account for comprehension of input from all modalities. Though much of the behavioral and neurocognitive evidence stems from experiments recruiting language or vision (Radvansky & Zacks, 2014), the theory advocates that its mechanisms are elemental procedures in human cognition[17] and yield a common representational format, rendering its tenets testable in creative experimental paradigms and across disciplines of cognitive science. In fact, the present thesis probes just this by comparing comprehension of events encoded in environmental sounds and spoken utterances. Therefore, this section focusses on a description of language comprehension according to event cognition.

By and large, Radvansky and Zacks' (2014) perspectives on language comprehension echo Zwaan's (Zwaan, Langston & Graesser, 1995; Zwaan & Radvansky, 1998) event-indexing model (see Chapter 2.4.2). Verbal, in contrast to visual input, for instance, naturally unfolds in a linear sequence and not all information is present at once. This compels comprehenders to construct series of working models (Zwaan's *current model*) in a piece-meal fashion. To understand verbal utterances, listeners segment the input and create small-scale sentential event models in acts of construal. In conversations, these event models are monitored for overlap with one another along situational dimensions, such as space, time, or agentivity, and hierarchically integrated into event models on a larger scale (Zwaan's *integrated model*). When dimensional overlap is absent, this integration fails and causes

[17] »the discourse-level comprehension mechanisms we have described here are not really about language as such, but about event cognition« (Radvansky & Zacks, 2014: 79; cf. also Zacks et al., 2007: 283) and the »proposal that event segmentation controls resource allocation and updates memory is a claim that event segmentation is a core, domain-general mechanism of cognitive control« (Zacks et al., 2007: 276).

prediction error, which signals the detection of an event boundary and triggers the building of a new working model while shifting the current event model into memory. Model construction from language is preceded by linguistic decoding processes that transform surface structures into propositions (Radvansky & Zacks, 2014: 57f.; van Dijk & Kintsch, 1983; Johnson-Laird, 1980).

### 2.4.4. Evaluation

Radvansky and Zacks' (2014) theory of event cognition is a more comprehensive theory of meaning construction than Kintsch's (1988) construction-integration model and Zwaan's (Zwaan & Radvansky, 1998) event-indexing model. Motivated by computationalism, Kintsch attempted to explain the construction of meaning from a strictly internalist perspective (Fodor, 1975). All knowledge inferences originate from bottom-up triggered spreading activations of propositional networks in the comprehender's mind. Coherence in discourse comes about through increasing activational overlap of nodes from sentence to sentence and expansions or reductions of the propositional network.

Zwaan, on the other hand, accounts for the fact that, to build coherent situation models, comprehenders track the dynamic unfolding of the events implied in the sequences of textual propositions. Zwaan and Radvansky (1998) laid the groundwork for Radvansky and Zacks' (2014) event cognition by rendering events the representational unit of comprehension, establishing the precursor to the prediction mechanism (monitoring of overlap), and considering influence of domain-independent features for coherence.

Event cognition theory, in addition to computational implementation (Reynolds, Zacks & Braver, 2007) and support from neurological evidence (Radvansky & Zacks, 2014: 52-56), expands the previous approaches by elaborating a model that holds for processing of input from all modalities, originating in the external and internal. It acknowledges top-down guiding of perceptual behavior by recruiting schematic representations of events in long-term memory. Event representations provide the common representational scaffold that underlies perceptual processing, action planning, memory encoding and recall, as well as language processing. As such, it embraces theoretical principles of embodied cognition about neurocognitive

architecture that assume distributed organization (Barsalou, 1999; 2008), which is the focus of the next chapter.

### 2.4.5. Summary

What happens in the mind when we comprehend input? Regarding the apprehension of both verbal and sensory input, Radvansky and Zacks (2014) presented the most sophisticated model so far. Adopting ideas from models of discourse comprehension (van Dijk & Kintsch, 1983; Kintsch, 1988; Zwaan & Radvansky, 1998), their theory of event cognition posits that the construction of meaning entails the (1) parsing up of verbal utterances or the perceptual stream into segments that are (2) conceptually integrated step-by-step into a structured mental representation. This process is aided by both (3) bottom-up and top-down flows of information and automated mechanisms that match the current input to stored knowledge.

Despite its elaboration, event cognition leaves important questions unanswered. It remains particularly vague about the precise nature of the matching process[18], i.e., how bottom-up input is compared to working models and how working models mesh with event schemata and vice versa, and it takes no definite stance on the representational format of event models. The following chapter therefore deals with an approach from cognitive neuroscience that formulated concrete hypotheses about the processes with which cognition taps into different knowledge repositories in the construction of meaning representations.

[18] Riemer (2015: 43) recognizes the risk of a regress argument: »If concepts are attributed to cognizers as part of the explanation for their intelligent action, then they also must have the capacity to apply the concept to situations. This capacity in turn requires us to attribute to them a whole new set of principles specifying the way in which this application happens, which themselves then require further principles for *their* own interpretation, and so on *ad infinitum*.«

## 2.5. Grounded cognition

In the second half of the 20th century, cognitive scientists sought alternatives to computationalist theories of cognition. Technological improvements of computers had inspired computer scientists to develop architectures of human cognition that could be implemented digitally on a machine (e.g., Minsky, 1974). Theorists dissatisfied with the premise that the mind is but an encapsulated, rule-based processor of abstract symbols (cf. Harnad, 1990; Searle, 1980) increasingly emphasized that cognition was a phenomenon that is also sensory-based and context-sensitive (e.g., Allport, 1985; Warrington & Shallice, 1984; Lakoff & Johnson, 1980). *Grounded cognition* (Barsalou, 1999; 2008; 2009; 2016; 2021) emerged as an influential framework because it has brought forward theoretical proposals as to how semantic memory is embedded in neuroanatomical structures, describing operational principles that underlie cognition, as well as how concepts are learned, stored, retrieved, and composed for various cognitive tasks.

In general, grounded cognition holds that mental representation arises from the activation of interconnected neural patterns across various brain modalities. This coordinated neural activity forms the physiological foundation of conceptual representation and enables us to process and interpret information in a meaningful way. What we experience as cognitive process (i.e., thinking, language, reasoning) is myriad cortical activity.

### 2.5.1. Storage of multimodal concepts

In now seminal work, Barsalou (1999) theorizes how humans encode, learn, and use perceptual information in cognition. Concepts (or *simulators*; Barsalou, 1999: 587) are learned when people attentively and repeatedly process components of perceptual experience. Importantly, perceptual experience is understood more broadly, including not only sensory but also stimulation from proprioception or introspection (Barsalou, 1999: 585). During repeated attentive processing, neural assemblies form in different brain areas and grow stronger in their association (Hebbian learning), imprinting themselves locally as neural units that represent experiential features and components. *Perceptual symbols*, by definition, are these »records of (…) neural activation« (Barsalou, 1999: 582). Whenever one is exposed to a particular experience

— be it through language, perception, or introspection — these previously entrenched neural patterns reactivate and »function symbolically, standing for referents in the world, and entering into symbol manipulation« (Barsalou, 1999: 578). Though easily misunderstood as conscious and holistic mental images of objects, perceptual symbols are rather like neurophysiological components that have encoded some information that is integral for building concepts.

Repeated co-activations of perceptual symbols in different modal systems associate them with one another and establish multimodal concepts in long-term memory. The cognitive outcome of these associations are categories, which are required when learning to distinguish cats from dogs, or animals from humans. To conceptually represent visual features of a DOG, e.g., neural patterns in the visual system fire that were encoded during experiences with DOG-exemplars. To represent barking, the auditory cortex becomes active, and so forth. As an aggregate, this distributed network of activity makes up the cortical correlate of the concept DOG (a *simulator*). Upon encountering a never-before-seen dog in the neighborhood, perceptual analysis dynamically reactivates and adapts the perceptual symbols in the sensory systems associated with the DOG-concept. The current sensory input interacts with stored experiential patterns and allows for the flexible recognition of that dog as a token of the DOG-category, enabling inferences about what might happen when one tries to pet it (Barsalou, 1999: 587). In other words, the same neural substrate supports perception *and* conceptual representation.

### 2.5.2. Simulation is conceptual representation

It follows that categorization of bottom-up input depends on reactivation of stored states of distributed neural activity. A network of activation patterns across multiple modalities instantiates the concept that is relevant for immediate processing. Barsalou termed this process of concept instantiation a *simulation* (Barsalou, 1999: 586f.; 2008, 2009, 2016). By definition, »(s)imulation is the reenactment of perceptual, motor, and introspective states acquired during experience with the world, body, and mind« (Barsalou, 2008: 618), serving the purpose »to represent information conceptually«

(Barsalou, 2021: 38).[19] According to grounded cognition, simulation is a fundamental and necessary principle of human cognition.

Although it is tempting to conceive of simulation as the apparition of mental images to the mind's eye, simulation refers to an unconscious reactivation of neural activity that, despite its effects on cognition (cf. Chs. 2.5.7, 2.6.3, 2.6.4), does not necessarily penetrate subjective experience as conscious internal visualizations or a vivid somatosensory reliving (Barsalou, 2016: 15; Tillas & Vosgerau, 2016: 466; Nanay, 2021). However, given that neuropsychological findings about mental imagery (Finke 1989; Kosslyn, 2005; Albers et al., 2013; Dijkstra et al., 2019) have repeatedly confirmed a functional involvement of the visual cortex in humans' ability to form mental images and reason with them, it is likely that mental images are the conscious, working memory representations of underlying, unconscious simulations in the visual system (Kent & Lamberts, 2008; for in-depth discussion, see Nanay, 2021). In this thesis, mental imagery is therefore understood as the phenomenological manifestation of an underlying, unconscious simulation. This is important because it not only profiles simulation as a basic mechanism to support cognitive processing but corroborates the argument that modality-specific brain areas support memory and conceptual representation in the absence of noticeable internal stimulation or direct sensory stimulation (Barsalou, 2008). In other words, not only are mental images the result of simulations, but simulation itself is an operation of conceptual representation.

This is precisely one of Barsalou's main tenets: simulation is a computational mechanism that underlies conceptual representation (2009: 1282). Various modal systems distributed across the cortex activate automatically to represent concepts from semantic memory. Neuronal units that have encoded perceptual symbols fire in an »entrenched associative network« (Barsalou, 2016: 14) and, as an aggregate, represent the integrated concept. When a concept is required, activation of its corresponding multimodal network, i.e., simulation, is automatic. Therefore, »grounded theories are often viewed as necessarily depending on (…) full-blown simulations that recreate experience« (Barsalou, 2008: 620). However, not all the perceptual symbols that make up a concept necessarily enter into conceptualization but may be activated selectively, or ad-hoc, to meet specific task demands (Barsalou, 2009: 1282; Barsalou, 2017; 2021). In fact, top-down signals from the prefrontal cortex can trigger simulations

[19] Allport (1985: 53) assumed that the »same neural elements that are involved in coding the sensory attributes of a (possibly unknown) object presented to eye or hand or ear also make up the elements of the auto-associated activity patterns that represent familiar object-concepts in 'semantic memory'«

selectively and target the specific modal and schematic components that manifest in conceptual representation. Mechelli and colleagues (2004) found that impulses originating in the prefrontal cortex[20] and terminating in sensory areas facilitate both perception and imagery of category-specific information, supporting the selectivity of simulations (see also Ishai et al., 2000).

Overall, simulation generates mental representations for various cognitive functions, including online and offline conceptual processing during, for instance, long-term memory recall, or recognition memory, where they represent previously encoded experiences (Kent & Lamberts, 2008; Laeng & Teodorescu, 2002). They also play a key role in constructive memory by allowing individuals to simulate hypothetical future scenarios based on past knowledge (Schacter & Addis, 2007). Section 2.5.5 below illustrates how simulation impacts language comprehension (for review, see Zwaan & Madden, 2005).

### 2.5.3. Event comprehension according to grounded cognition

In later work, Barsalou (2009; 2016) developed theoretical constructs that underlie efficient online processing of events. *Situated conceptualization* and *pattern completion inference* drive our understanding when information from perception, action, and cognition converges, e.g., when somebody engages in natural activity and processes complex sequences of input (≈ *situated action*; Barsalou, 1999). This section describes how situated conceptualizations and pattern completion inference support event comprehension.

Events are made up of different components, each of which might activate distinguishable concepts. Stair-climbing events, for instance, involve at least a figure, the motion of that figure, and a ground where this motion takes place (Talmy, 2000a; 2000b; see Chapter 2.6). Of course, »(r)ather than perceiving elements of the situation individually, they are experienced globally as a coherent conscious state« (Barsalou, 2016: 16). Categorization of sensory input and clustering of activated conceptual components yield structured working memory representations called *situated conceptualizations*. Situated conceptualizations are referential mental models[21] and,

[20] The prefrontal cortex supports cognitive functions such as attention, inhibitory control, and memory (Miller & Cohen, 2001; Friedman & Robbins, 2022) and is thus crucial for the selective activations of simulations.

[21] Like Johnson-Laird's (1980) mental models, Radvansky and Zacks' (2014) event models, or the situation models proposed by Zwaan and colleagues (1995), as discussed in the previous chapter.

as such, can be encoded into long-term memory as episodic exemplars (e.g., that one time you saw two people riding a tandem bicycle) or as instantiations of structured schemata (categories or types of situations, e.g., bike-riding) (Barsalou, 2009: 1283).

In online processing, situated conceptualizations arise in working memory as a product of simulation. While listening to an anecdote about a bike-riding event, for instance, perceptual symbols associated with bike-riding events fire to represent concepts from long-term memory that are integrated into coherent representations. Importantly, coherence of these situated conceptualizations is constrained by both the idiosyncrasies of the bottom-up input – here, the construal explicit from verbal utterance structure – and the schematics of top-down semantic representations.

Repeated exposure and statistical learning cause that »(c)omponents of the conceptualization become entrenched as simulations in the respective simulators, as do associations between simulations and simulators« (Barsalou, 2009: 1284). In other words, a concept can become associated statistically with certain situated conceptualizations and with potential simulations. Thus, situated conceptualizations may become schematic long-term memory representations, providing a knowledge store that facilitates mental model construction. Because of this, they enable *pattern completion inference*, which is Barsalou's (2016: 19) term for prediction, in online processing of input. The crucial mechanism underlying this sort of predictive inference is, in accordance with grounded cognition, multimodal simulation. For example, one can simulate, or predict, the approximate trajectory of a pencil that is rolling towards the edge of a table and about to fall to the ground. Because situated conceptualizations make that predictable, a reactive movement of the hand to a certain location allows one to intercept the pencil; they allow agents to engage in situated action: the schematic information activated in situated conceptualizations, here a linear trajectory in space, prepared the pen-catcher for catching. This information is encoded in perceptual symbols, which, as detailed in section 2.5.2, are reactivated through simulation. From a neurocognitive perspective, multimodal simulations are a plausible and effective representational format for prediction:

> »Because simulated predictions reside in the same systems that perceive the environment, carry out actions, and introspect on internal states, they can be matched to actual experience as it occurs, thereby assessing whether events have unfolded as predicted« (Barsalou, 2009: 1284).

Consequently, simulation can account for the mechanism in Radvansky and Zacks' (2014) event cognition theory that enables the matching of top-down event schemata with the event segments extracted from the bottom-up perceptual stream (cf. Chapter 2.4.3). The blending of a top-down stream with the bottom-up stream, that is, semantic representation with sensory representation, comes naturally in grounded cognition because both are encoded by perceptual symbols and use the same brain infrastructure.

### 2.5.4. Grounded cognition complements event cognition theory

Overall, many of Barsalou's theoretical assumptions are compatible with tenets of Radvansky and Zacks' (2014) theory of event cognition and complement it in two important respects.

Both theories are representationalist and grounded, assuming multimodal conceptual representations that underlie comprehension. Working models are established by a matching of bottom-up input with top-down concepts that are associated with structured schemata. Grounded cognition is more specific concerning the mechanism with which event models consolidate, namely through simulation (generating *situated conceptualizations*, a conception) of schematic information contained in long-term memory representations (a concept or simulator). Simulators can incorporate event schemata, that is, event models that have become entrenched through repeated experience or habituation. Activation of these top-down knowledge representations allows for prediction (Radvansky & Zacks, 2014) or, respectively, pattern-completion inference (Barsalou, 2009).

Nevertheless, Barsalou's theory is more explicit than Radvansky and Zacks' (2014) regarding the prediction mechanism (Barsalou, 2016: 19). Since simulations based on top-down schemata reside in the same modality-specific neural systems as perceptual representations of the actual experience (bottom-up input), matching occurs as the input unfolds.

### 2.5.5. Perceptual simulation in language processing

Arguing that perceptual simulation underlies conceptual processing challenges a strict separation of amodal and modal representations, as both higher-level (inference,

planning, language) and lower-level functions (perception, motor control, memory encoding) equally depend on modal simulations. Understanding a sentence relies fundamentally rely on the activation of attributes associated with referenced entities or details pertaining to described events. Grounded cognition thus aligns with models that propose distributed cognitive architecture (cf. Singer, 2009). From its core thesis that simulation is the main computational principle in this distributed system follows the assumption that representations of verbal utterance meaning are eventually the same as those for the interpretation of the perceptual experience referred to in that utterance. Moreover, simulation of concepts is indifferent to input modality, meaning it does not matter whether verbal, sensory, or introspective input activated a concept because the activation sets off a multimodal simulation tailored the current task, providing a mental representation of meaning. This foundational principle of Barsalou's (2008) theory is critical to the present study, anchoring its research question and motivating its hypotheses (cf. Ch. 2.7): representing meaning for environmental sound or speech input, as well as representing meaning for language comprehension and language production, all relies on the core operation that is simulation.

That said, how would one comprehend the sentence 'a car passed by'? Analysis of verbal input (see Ch. 2.1.1) activates concepts and relates them to one another in a construal (Langacker, 1986; Zwaan & Madden, 2005; Radvansky & Zacks, 2014) derived from morphosyntactic structure. Respecting the hierarchical configuration of the entities in a construal, parallel simulation of the relevant concepts integrates perceptual symbols into a situated conceptualization — resulting in an event model. This way, the verbal input constrains comprehenders' simulations with respect to selectivity, focus, or perspective (Zwaan & Madden, 2005: 233). The situated conceptualization profiles the most relevant sets of components of CAR and PASSING that are required for comprehension and backgrounds currently irrelevant subordinate concepts like TRUNK or ENGINE (Lupyan, 2012; Langacker, 1986).

At the same time, a situated conceptualization is not restricted to the lexical core concepts in the utterance but steadily enriched with schematic attributes of the respective category. This may include simulations of schematic knowledge, such as typical locations where cars engage in passing-by, spatial or causal relationships between the car and the entity that was passed by, the temporal extension of the conceived event, etc. When the input is language, the referenced event is not necessarily available for sensory inspection. The situatedness of conceptualizations,

however, gives comprehenders the option of an *as-if* modal processing due to the underlying neuronal activity »creating the experience of 'being there'« (Barsalou, 2009: 1283) (see also Fuchs, 2018: 200f).

### 2.5.6. Language as an independent representational medium

In a series of experiments, Barsalou and colleagues (Solomon & Barsalou, 2004; Simmons et al., 2008; Barsalou et al., 2008; Santos et al., 2011) found that simulations are not the only form of representation underlying conceptual processing and made a case for language as an independent form of representation (cf. Paivio, 1971).

Simmons and colleagues (2008) used functional magnetic resonance imaging (fMRI) to compare the cortical activity of ten subjects engaged in property generation tasks. In a first session, participants were asked to come up silently with conceptual features of 30 stimulus words randomized across 5 blocks: »What characteristics are typically true of X?« (Simmons et al., 2008: 110). Each trial lasted 15 seconds. In a later localizer session, word association and situation generation tasks were administered, with the former demanding conceptual processing based on language and the latter on simulations[22], revealing participants' most active brain areas during lexical processing or mental imagery. The imaging data from the property generation task, when participants were not explicitly instructed to use language or situated simulation, was compared to the data of the localizer session to study the loci of cortical activity during conceptualization[23]. Results showed consistently that, during the first half of each 15 second trial, brain areas responsible for lexical processing exhibited stronger activity than those responsible for producing mental images. This pattern reversed in the trials' second half, where blood flow into language areas decreased in favor of stronger activity in areas dedicated to situated simulation (Simmons et al., 2008: 115). Simmons and colleagues (2008) interpret this as revealing a time course of conceptualization of meaning. Upon encountering a concept in verbal form, verbal representations initially dominate conceptualization because activity peaks in language areas of the brain first. Once »the language system runs out of responses, however, attention may shift to the simulation system« (Simmons et al., 2008: 116), as

[22] »For the following word, what other words come to mind immediately? […] and […] imagine a situation that contains what the word means and then describe it?« (Simmons et al., 2008: 111)
[23] This is not how Levelt (1989) used the term, but Simmons et al.'s (2008) use highlights the integrative quality of conceptual representation and is henceforth used synonymously to conceptual representation.

suggested by increased peaking of cortical activity in distributed modal regions during the second half of each trial.

These findings indicate that in response to linguistic input, people tap into simulation systems only after they have consulted language systems. This suggests, first, that verbal decoding processes precede simulations of verbal concepts. Second, it aligns with the principle of encoding specificity (Tulving & Thomson, 1973), in that »information in memory most similar to the cue becomes active most rapidly« (Barsalou et al., 2008: 249). Third, representational activity in language areas might suffice to manage comprehension (Ferreira, Bailey & Ferraro, 2002), as the activity in sensory systems is delayed and might succeed a semantic decision. Fourth, response time analyses from sensory-based decision tasks have shown that duration of simulation is negatively correlated with concreteness of verbal stimuli (Gerwien et al., 2024), suggesting that modal simulations are the more demanding way to represent conceptual features when words are less imaginable. Moreover, additional neuroimaging evidence suggests that simulation may not be mandatory in language comprehension, but task- or stimulus-dependent instead (Just et al., 2004) and possibly only integral when deep conceptual processing is required, like during property generation (Simmons et al., 2008; Handjaras et al., 2015), narrative or discourse understanding, and not in situations where a shallow comprehension is enough to get by, like in lexical decision tasks or structural priming (Kemmerer, 2015). Altogether, converging findings contradict Barsalou's (1999) radical stance of indispensable modal simulation in conceptual processing and attest that meaning, at least of verbal utterances, may be processed without causal contributions of simulations.

Arguments like these motivated Barsalou and colleagues (2008) to formulate the *language and situated simulation (LASS)* theory. Its main tenet is that both linguistic forms and multimodal, context-sensitive simulations are main representational media to enable conceptual processing (cf. Truman & Kutas, 2024; Kemmerer, 2015). Importantly, linguistic forms are not understood as amodal symbols but lexical items that are statistically associated in natural language (like in linguistic corpora or large language models), and consequently cause the underlying concepts to become associated with one another as well (cf. Dove, 2022).

In LASS, conceptual processing is hence modeled as a continuous interaction of verbally triggered concepts and situated simulations. Although linguistic forms (cf.

Chapter 2.1) and situated simulations set on quickly after exposure[24], linguistic processing dominates initially. Words activate other, statistically associated lexemes, which in turn index concepts (like *labels*; Lupyan, 2012: 4) that are connected to the one currently processed (like in the network structures assumed in connectionist models, cf. Kintsch, 1988). At this point, the conceptual representation is built mainly from concepts connected to words that appear in syntactic structures. When superficial processing of linguistic forms is not sufficient to satisfy comprehension, activation »of deeper conceptual information is necessary« (Barsalou et al., 2008: 249) and the conceptual system turns to the previously unattended, yet concurrent situated simulations to continue conceptual representation.

Segmenting conceptual representation into different phases or modes is reminiscent of Kintsch's (1988) construction-integration model, with the construction phase constituting as activation of lexeme nodes and the integration phase conceived of as a delving into situated simulation with the goal of situation model construction (Barsalou et al., 2008: 268). Linguistic processing directly supports the building of meaning representations in language comprehension in that »words and syntactic structures function as cues to assemble a simulation compositionally« (Barsalou et al., 2008: 252). In other words, the unfolding verbal input guides construal (Langacker, 1986; Lupyan & Bergen, 2016: 415; Bocanegra et al., 2022), which resonates grounded cognition's fundamental assumption that meaning is represented by simulations of concepts, and that verbal labels activate these concepts and structure the simulations (Lupyan, 2012).

### 2.5.7. Evidence for perceptual simulation

Among the myriad hypotheses brought forward by grounded cognition, the simulation mechanism has been the strongest impetus for experimental research in different fields. How a neural reenactment of multimodal representations affects cognitive processing has been demonstrated in a variety of studies. Simulation is reported to exert solid effects on behavioral measures in conceptual tasks (e.g., in mental scanning tasks of Kosslyn et al., 1978; sentence-to-picture-matching in Stanfield & Zwaan, 2001; in visual detection tasks by Craver-Lemley & Arterberry, 2001; for a

[24] Amsel and colleagues (2014) reported that manipulations of a visual stimulus took full effect on semantic processing of word meanings at already 200 ms after stimulus onset. This is compatible with findings by Pulvermüller (1999).

review of the behavioral evidence, see Dijkstra & Post, 2015) or on peripheral physiological activity like muscular impulses (e.g., in facial electromyography in response to verbs, see Foroni, 2015), eye movement (see Chapter 2.6.4), or on the strength of event-related potentials in different cortical regions (Amsel et al., 2014). Crucially, the functional involvement of simulation in conceptual representation is evidenced by selective disruptions of central physiological activity in, e.g., the motor cortex, via transcranial magnetic stimulation (TMS) (Pulvermüller et al., 2005; Buccino et al., 2005; see Bonner & Grossmann (2012) for lesion fMRI evidence; see Noppeney (2009) for a systematic review of neuroimaging evidence).

In a seminal study, Kosslyn and colleagues (1978) conducted multiple experiments in which participants studied visual stimulus arrays or images and were subsequently asked to perform different judgments or manipulations of the stimulus images in memory. Critical conditions modified spatial information in the stimuli, such as distances between critical items, number of present distractor stimuli, as well as the size of the stimulus images. In each experiment, participants were explicitly instructed to encode the stimuli as mental images. After a short study phase, the stimulus was covered, and participants had to solve tasks about visuo-spatial qualities of the previously shown critical items. Analyses confirmed that response times were significantly correlated with the distances between the prompted critical items. It took participants engaged in visual imagery more time to scan across images when the distances to be scanned are larger compared to smaller distances. This constitutes evidence for grounded cognition in that, for one, the working memory representations underlying visuo-spatial reasoning are perceptual analogues, and second, that the processing mechanisms to scan these internal visual representations are commensurable with those used for perception of external stimuli. Altogether, Kosslyn and colleagues' (1978) study suggests common mechanisms in perception and cognition and supports the argument that mental representation relies on activity in the sensory modalities.

Stanfield and Zwaan (2001) report psycholinguistic evidence from a sentence-to-picture-matching task. Participants pressed a button when they recognized the display of a colored drawing as an object that was mentioned just before in a written stimulus sentence. An automatic simulation of sentence meaning, so their prediction, would bias the visual system to anticipate the stimulus object image in a certain spatial

orientation and thus speed up recognition. For instance, simulations based on sentences like »John put the pencil in the drawer (vs.) […] in the cup« (Stanfield & Zwaan, 2001: 154) should cause faster recognition of, respectively, a horizontally-oriented or vertically-oriented image of a pencil. Response time distributions confirmed effects of the orientation condition. If the object image congrued with the inferred orientation, response times were faster than in the incongruent condition. This implies that meaning representations of verbally encoded events are perceptual simulations that comply with context-specific construals (Barsalou, 2009).

While Stanfield and Zwaan (2001) concluded facilitatory effects of simulations, Craver-Lemley and Arterberry (2001) found the opposite. Motivated to elucidate whether internal visualizations interfere with or support visual perception, they conducted a series of experiments with a visual detection paradigm. Participants were instructed to map vertical or horizontal bars, which they had been shown in the beginning of the experiment, from memory onto a grid of five rectangular cells shaped like a plus sign. Importantly, they were asked to maintain fixation on the center cell. The task required detection of a rapidly presented asterisk (ø = 16 ms) in the top or bottom cell in critical trials. Detection performance was impeded when participants were occupied with mapping an internal mental image onto an external visual percept. If attention is needed to maintain simulated visualizations top-down while blending them with a visual percept streaming from bottom-up, this process seems to bind resources to a degree that reduces sensitivity of the visual system to external stimulation (Lavie, 2005). This interference suggests, at least, that cognitive and perceptual processes compete for one and the same representational substrate.

Importantly, Craver-Lemley and Arterberry's (2001) findings mainly illustrate how within the visual modality, low-level attention to an external percept is impeded when internal representations occupy the visual system. Stanfield and Zwaan's (2001) task, in comparison, required a sequential crossing of modal representations, with visual working memory receiving input from verbally induced simulations first, and subsequently using the simulation, like a prime, for stimulus recognition. Differences aside, both studies demonstrate how simulations, as concerted activity of neural circuits for perception, affect attention and cognition in online tasks.

Despite such convincing behavioral effects, the cited evidence is indirect and does not confirm a functional dependency of conceptual representation on neural activity in

modal systems (Hauk, 2016). To test such causal involvement, various studies have employed transcranial magnetic stimulation (TMS). TMS is technology that allows locally specific disruptions or boosts of cortical activity.

Pulvermüller and colleagues (2005) have shown that response times in a lexical decision task were significantly shorter for verbs labeling hand- or foot-actions when concordant hand- or foot areas of the motor cortex were stimulated than when no stimulation was applied. Increasing the resting-state activity of specific neural assemblies with TMS sped up participants word processing and yielded a faster decision. This strongly suggests that these motor areas become functionally engaged for the semantic processing of verbal input.

In an elegant study, Buccino and colleagues (2005) similarly applied TMS to the motor strip and measured electromyographic potentials at the hands and feet when participants comprehended sentences that selectively expressed actions performed with these extremities. Two interesting findings emerged. First, both electromyographic and response time measures were negatively affected when the verbal input referred to the corresponding hand or foot action. This suggests that interpreting action-related sentences interfered with motor cortex excitability through TMS because semantic processing occupies local cortical resources. Already engaged in motor simulation to represent sentence meaning, the motor cortex exhibits reduced responsiveness to the TMS impulse, attenuating the signal that was ultimately measurable at the effector muscle[25]. In a second experiment, response latencies were measured for participants' cognitive judgements about whether the sentences expressed hand- or foot-related actions. In line with the interference found in the TMS data, the response time data obtained from button presses with hand or foot depended on the actions referred to in the verbal stimuli. If hand-actions were expressed, responses given by button presses with the hand were slowed down, and foot data correspondingly so. The research by Buccino and colleagues (2005) and Pulvermüller and colleagues (2005) indicates that understanding of verbally encoded actions is grounded in sensorimotor systems.

[25] Importantly, Pulvermüller et al. (2005: 796) applied TMS at subthreshold signal strength (~90% of the excitability threshold), whereas Buccino et al. (2005: 357) used stronger intensity of 120%. The excitability threshold refers to the minimum level of TMS stimulus intensity at which motor-evoked responses can be reliably measured from an individual's effector muscles (Rossini et al., 1994). Moreover, in Pulvermüller et al.'s (2005) lexical decision task, the TMS impulse fired at 150ms after word onset, while in Buccino et al. (2005), stimulation was delivered depending on length of the verb root, by and large between 500-700 ms after sentence onset.

Targeted use of TMS to tamper with this neural substrate modulates human subjects' performance in language-related tasks (cf. Shapiro et al., 2001).

Somatic effects of simulation on the peripheral nervous system in online language processing have further been reported by Foroni (2015). In a cross-linguistic study of written sentence processing, electromyographic potentials were measured from the facial muscle group that enables smiling. This activity was increased in both the native and foreign language for phrases about actions depending on that muscle group (e.g., 'I am smiling'), notably already within 200 ms of verb offset (Foroni, 2015; Foroni & Semin, 2013). Surprisingly, the muscle impulse was weaker in the foreign language, which suggests that the bodily grounding of verbal concepts, i.e., effects of conceptual processing on the peripheral nervous system, is mediated by language proficiency or immersed experience. This transpires to be a compelling argument for LASS theory (Barsalou et al., 2008). Language, as a symbol system, is an inventory of symbolic labels for grounded endogenous and exogenous experience (word-level grounding), and at the same time, it is a medium for conceptual representation that is itself acquired as grounded in experience (language grounding) (cf. Dove, 2022). Attenuated grounding of a foreign language (Pavlenko, 2012) illustrates that language may represent concepts without sufficient contributions of simulations, hinting at a division-of-labor of language and situated simulation in conceptual processing (cf. shallow vs. deep processing; Barsalou et al., 2008; Zwaan, 2016).

### 2.5.8. Summary

In summary, grounded cognition posits that human cognition does not occur within a specialized, isolated module dedicated to processing amodal concepts. Instead, it is neurally and conceptually grounded in the brain's modal systems responsible for perception (e.g., vision, audition, somatosensation) and action (proprioception, motor control) (Barsalou, 2008). Knowledge of concepts is not stored exclusively in an amodal format — such as the *language-of-thought* proposed by Fodor (1975) — but rather as *simulators* (i.e., concepts) that cluster perceptual symbols (Barsalou, 1999; see Ch. 2.5.1). A simulator is established through the concerted activity of neural units that have encoded perceptual symbols of referent properties. Conceptual representation is achieved with a simulation mechanism, which is defined as the reinstantiation of associated patterns of neural activity that would occur during

immediate perceptual experience of a situation, verbal reference to it, or memory and introspection thereof. Systematic examinations of activity in the central and peripheral nervous system, as well as behavioral findings support these assumptions.

With *language and situated simulation* (LASS), Barsalou and colleagues (2008) advanced a theory that is detailed enough to enable experimental testing of grounded cognition hypotheses, for instance about time course and extent of simulation in conceptual processing for language comprehension. A central prediction in LASS is that processes of language decoding run in parallel to situated simulation, with the conceptual system turning to one or the other in response to task demands, biasing conceptualization towards a language or situated simulation mode. Concurrently to initial focus on a language mode for processing of verbal input, situated simulations are generated 'behind the scenes' and held available for efficient tapping into conceptual depth. If these unconscious simulations — much like language decoding processes — are integral to language comprehension from the initial exposure to verbal input, then simulation-related neural activity in modal systems can be expected to influence oculomotor patterns during language comprehension. It is one goal of this study to test this assumption.

Accordingly, experiments were designed to analyze eye movement patterns during the processing of motion events, which are presented either as non-verbal environmental sounds or as verbal utterances. The following chapter describes the conceptual basis of motion events (Talmy, 2000a; 2000b) and reviews previous eye-tracking research on simulation in motion event processing.

## 2.6. Motion event conceptualization

Inspired by the late 1960s search for linguistic universals (Greenberg, 1966), Talmy dedicated himself to describing a wide range of languages, notably those spoken by Native American groups (Talmy, 1972), focusing on systematic analyses of linguistic structures used in the expression of events. His major contribution to the field of cognitive science concerns how we conceptualize space in motion events when we describe them through language. Talmy's cross-linguistic analyses revealed typological differences between languages in form-meaning mapping (*lexicalization patterns*), that is, which conceptual components of motion events are selected for overt lexicalization across surface structures (Talmy, 2000b: 21). Some languages typically lexicalize information about a path of motion in the root of the predicate verb (verb-framed languages), while others allocate this information to an adjunct to the verb (satellite-framed languages) and instead express another quality of the motion event (e.g., manner, cause). Simply put, when speakers of different languages speak about motion events, the linguistic surface structures available to them vary with respect to the (objective) motion situation's features they encode.

### 2.6.1. The conceptual building blocks of motion events

Preceding Radvansky and Zacks (2014), Talmy defined *event* similarly: an event is a portion of the perceptual stream that is delineated by boundaries (2000a: 261). Within these boundaries, there are distinguishable elements, which are possibly universals of visuo-spatial perception (Talmy, 2000a: 220) and used to structure cognitive representations of events (Talmy, 2000b: 215) across modalities (Talmy, 2000a: 260).

Meticulous examinations of language examples (Talmy, 2000b) led to the assertion that motion events, defined as situations »containing motion and the continuation of a stationary situation alike (…)« (Talmy, 2000b: 25), consist of components that fulfill specific roles in the construction of meaning (Talmy, 2000b: 214ff.). First and foremost, there are *figure* and *ground*, notions adopted from Gestalt theory in perceptual psychology (Wertheimer, 1922/2017). The figure refers to »a moving or conceptually movable entity« (Talmy, 2000a: 184) and »generally the component on which attention« is focused (Talmy, 2000b: 218). The ground, on the other hand, 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, 2000a: 184). The *activating process* is an operation that, mostly encoded in the predicate verb, brings about a dynamic construal of an event, relating the figure to an instance of translational motion or a remaining stationary. The *association function* specifies the relationship between the figure and its ground and thus constitutes the motion event's *core schema*, such as the path that a moving figure takes with respect to the ground (Talmy, 2000b: 218). For example:

<table>
<tr><td colspan="2">The horse</td><td>raced</td><td>past</td><td colspan="2">the barn.</td></tr>
<tr><td>ART.DEF</td><td>NP</td><td>V.3sg.</td><td>PREP</td><td>ART.DEF</td><td>NP</td></tr>
<tr><td colspan="2" rowspan="3">figure</td><td>MOVE+MANNER</td><td>PATH-SATELLITE</td><td colspan="2" rowspan="2">ground</td></tr>
<tr><td>activating process</td><td>association function</td></tr>
<tr><td></td><td colspan="3">core schema</td></tr>
</table>

***Table 2-2:*** Conceptual components in motion event representations according to Talmy (2000b).

Importantly, in Talmy's theory, event representations contain a »privileged core« (Talmy, 2000a: 260) that constitutes the essence of an event: »language users apparently tend to conceive certain elements and their interrelations as belonging together as the central identifying core of a particular event or event type« (Talmy, 2000a: 259). To legitimize the psychological reality of this core, he reviews a case study of a deaf child whose event-referring sign language suggested an underlying conceptual structure that is commensurable with the one structuring verbal expression (Talmy, 2000a: 300-303).

### 2.6.2. The importance of the path-component

There are multiple arguments to support that the trajectory of motion (≈ path) is the conceptual component that constitutes this core in translational motion events. First, Talmy's typology is centered on where on the surface languages encode path, yielding the verb-framed versus satellite-framed distinction (2000b: 221-222). Second, path uniquely specifies the trajectory or direction of motion, rendering it both necessary and sufficient for a situation to be conceived of as a translational motion event. This ascribes to path a crucial role in motion event conceptualization (2000a: 265ff.; 2000b: 227) and aligns with the notion that path is the information encoded in the *core schema* (*association function* + *ground*) of the motion event (Talmy, 2000b: 222). Third, as a

perceptually primitive component, path can be sensed across modalities[26] and drawn on in the »structuring in other cognitive systems such as visual perception« (Talmy, 2000a: 260; Marr, 1982). The path component in motion event schemata is considered a qualitatively »idealized path schema« (Talmy, 2000a: 149) that comprehenders apply to referent situations (Talmy, 2000a: 220f.). When understanding language or construing a producible message, idealized path schemata provide structure to conceptual representations of motion events, much like image schemas (Johnson, 1987), idealized cognitive models (Lakoff, 1987), or spatial primitives (cf. Chatterjee et al., 1999). Path schemata have their conceptual basis in topology, specifically in abstract geometric templates (Talmy, 2000a: 165f; cf. Herskovits, 1986; Hochberg & Fallon, 1976). For instance, in a figure's traversal *along* a trajectory (Talmy, 2000a: 185ff.), the vector of motion is conceived of as an abstract straight line mapped in relation to a ground and the figure as a point moving along that linear vector (Talmy, 2000b: 53; Talmy, 2000a: 106ff.). Note that the preposition *along* is the crucial surface element to express the figure's schematic motion trajectory of the figure. Accordingly, Talmy notes that a »main characteristic of language's spatial system is that it imposes a fixed form of structure on virtually every spatial scene« (Talmy, 2000a: 181; cf. Landau, 2017).

The present project agrees with the privileged role of path in motion event conceptualization by viewing the idealized geometric paths as the crucial component to provide structural coherence in motion event models (Radvansky & Zacks, 2014). Path schemas are essential for identification of a token-situation as a member of a translational motion event-type, telling us what the event is and how it unfolds (cf. Slobin, 2004: 238). If the core schema components (i.e., transition on ground) are necessary for the conceptualization of motion events, you cannot think of a translational motion event without thinking at the same time of a figure entity moving through space. This has important implications for the formulation of the hypotheses (cf. Ch. 3.1): perceptual simulations of the path components activated from motion event schemata are assumed to induce systematic eye movement behavior.

[26] It can be *seen* by maintaining gaze on a figure on a spatial trajectory, it can be *heard* as the sounds emitted by the motions of that figure, it can be felt *proprioceptively* by moving our bodies and *motorically* by enacting gestures. This holds equally for the quality of manner.

***Summary of Talmy's main ideas***

So far, this section described the central components incorporated in meaning representations of translational motion events according to Talmy (2000a; 2000b). Motion events are a fundamental experience that everybody can relate to, and everybody has schemata for them. Talmy has shown that the conceptual building blocks with which motion event representations are built – figure, path and ground – are shared across many languages (cf. von Stutterheim, 2017: 49). On top of that, motion events make convenient stimuli because figure, path and ground are grounded in visual (sensory) perception of space (Talmy, 2000a: 91; 162). This makes them not only adequate objects of investigation in typological research, but for the cross-modal comparisons in the eye-tracking experiments of the present study.

### 2.6.3. Evidence for motion event schemata

Talmy's extensive work has been the impetus for myriad empirical investigations of motion events in many different languages (von Stutterheim et al., 2012; Gerwien & von Stutterheim, 2018; Papafragou, Massey & Gleitman, 2006; Flecken, Athanasopoulos, et al., 2015; Flecken, Carroll, et al., 2015; Zlatev & Yangklang, 2004; von Stutterheim et al., 2020; Papafragou, Hulbert & Trueswell, 2008; Bohnemeyer & Pederson, 2010; Habel & von Stutterheim, 2000; Berman & Slobin, 1994), which, in turn, inspired theoretical amendments (Slobin, 2004; Pourcel, 2005; Carroll et al., 2012). The evidence attests that speakers of different languages habitually encode different configurations of Talmy's components and that their attention distribution on these components varies (Talmy, 2000b; Slobin, 2004). A valid conclusion from the evidence, which is often taken for granted, is that the components become relevant in the conceptualization of motion events one way or another. It may seem trivial, but the purpose of this subsection is to review evidence supporting that conceptual representation of motion events depends on building motion event representations with Talmy's schematic conceptual components, as stored in event schemata (Radvansky & Zacks, 2014).

Gerwien and von Stutterheim (2018) conducted a cross-linguistic study with native speakers of German and French and tested event segmentation in non-verbal and verbal experiments. Looking at short video clips of translatory motion events, one

sample was instructed to verbalize what is happening in the video (verbal condition), while another was tasked to press a button whenever something new occurred in the videos (non-verbal condition). Since French, a verb-framed language, requires its speakers to express path-information in the main verb and does not allow clusters of multiple path-segments in a single predicate, the authors predicted that speakers of French would (a) perceive new event segments and (b) produce a new verbal predicate whenever qualitative changes in the figure's trajectory occurred (e.g., changes in direction or orientation). In German, incorporation of multiple path adjuncts into the verbal predicate is possible, as the semantic element expressed is a property of the figure object (manner of motion) that remains the same across these multiple path segments, thus not demanding the conceptualization of a new motion event based on path-changes. Accordingly, the authors expected the stimulus events to be verbalized as a single assertion and fewer button-presses compared to the French sample. Results showed that, indeed, French speakers produced significantly more assertions on average than German speakers and were significantly more likely to perceive event breakpoints at path-changes. This indicates an effect of language on perceptual processing, a cognitive operation traditionally assumed to be non-verbal. Nevertheless, the German participants, despite habitually attending to and expressing manner in the conceptualization of motion events (Carroll et al., 2012; Talmy, 2000b), detected new event segments at path changes as well (Gerwien & von Stutterheim, 2018: 232), suggesting that they are, in fact, sensitive to the path component in a non-verbal task, though to a lesser degree than the French. Gerwien and von Stutterheim (2018: 234-235) suggest that these language-specific preferences in the non-verbal task arise from an influence of habitual linguistic expression on the weighting of attributes in event schemata. While a general motion event schema contains, i.a. *direction*, *orientation* and *manner* components[27], language experience may cause these abstract event schemata to become profiled for components that are focused by the grammaticized means of a language to express these events. Such profiled event schemata supply the building blocks for the working model (see Chapter 2.4.3, Radvansky & Zacks, 2014), which Gerwien and von Stutterheim (2018) consider the

[27] These more precise dimensions contrast with the coarse *path* and *ground* components assumed by Talmy (2000b). In Talmy's framework, *orientation* counts as a feature exhibited by the *figure*, whereas *direction* is fused in *figure* and *path*; both are omnipresent and objective qualities in translational motion events but not regarded within Talmy's taxonomy and, hence, not considered crucial for event representation.

conceptual representation that underlies both message generation in language production, as well as the button-press decision in event segmentation.

Evidence from a previous cross-linguistic study by von Stutterheim and colleagues (2012) supports similar language-specific effects on motion event conceptual representation. Analyses of gaze data, verbalizations, and memory performance all confirmed that speakers of languages with grammaticized aspect allocated more visual attention to, verbalized, and memorized the event as ongoing (e.g., focusing on the truck *driving on a country road*), whereas speakers of languages without aspect interpreted the stimuli in a holistic way, including on potential endpoints of the depicted motion events (e.g., *ein Lastauto fährt zu einem Dorf*, in Engl., a truck *is driving towards* a small town). Interestingly, speakers of non-aspect languages exhibited better recall of the endpoint-objects than speakers of aspect languages, suggesting effective structural differences in components of the event models that are encoded into memory, as well as what was salient during event perception. In the discussion, von Stutterheim and colleagues (2012: 859) note that the English-speaking sample, although equipped with progressive aspect, increasingly fixated the endpoint-objects during articulation, contrasting with the other speakers of the aspect-language population (cf. von Stutterheim et al., 2022: 31). This observation implies that the language-specific event schemata applied in event apprehension (Gerwien & von Stutterheim, 2018) do not rigidly determine the components in the emerging event model representation, but rather bias the working model underlying message generation in language production (von Stutterheim et al., 2012: 862).[28]

The reviewed evidence shows that speakers of different languages profile different conceptual components in online (Gerwien & von Stutterheim, 2018; von Stutterheim et al., 2012) and offline (von Stutterheim et al., 2012) representations of motion events. However, despite the lack of surface forms available for verbal expression, they all have the means, cognitively and perceptually, to conceive of motion events in non-habitualized ways. These means derive from the conceptual categories described in Talmy's framework: *figure*, *ground*, the relational *activating process* and *core schema* (= path). Handling these components and manipulating the entrenched configurations allows for typologically atypical event construal.

[28] On another note, von Stutterheim and colleagues' (2012) findings suggest that conceptualization of motion events does not hinge exclusively on the dimension of space but the dimension of time, as aspect modifies the configuration of event segments in terms of temporal clustering.

Beyond the typological debate, this review demonstrates, at least, that people gain an understanding of motion events by drawing on the conceptual categories described by Talmy (2000a; 2000b) from motion event schemata (Gerwien & von Stutterheim, 2018; Radvansky & Zacks, 2014)[29]. Simply put, and considering the global scheme of this study, people think about figures moving through space when they comprehend motion events.

### 2.6.4. Evidence for simulation during motion event conceptualization

As described, event schemata assume a fundamental role in yielding structure to event representations in tasks of verbal and non-verbal comprehension. Laying the empirical substrate for the research questions tackled in this project, this section discusses experimental findings indicate mental simulation of the components stored in event schemata (Barsalou, 2008; see Chapter 2.5.3). The evidence presented is derived mainly from response time measurements and analyses of eye-tracking data.

#### 2.6.4.1. Evidence from response time measures

The studies reviewed in this section interpret their findings on the basis of so-called Perky effects (named after Perky, 1910; see Craver-Lemley & Reeves, 1992). The Perky effect is assumed to occur when one's internal visual stimulation, i.e., a mental image, interferes with the detection or discrimination of an external visual stimulus (Richardson et al., 2003: 770).

Richardson and colleagues (2003) conducted two experiments. In experiment 1, they expected Perky effects in the detection and recognition of shortly presented shapes in certain positions on the stimulus screen. After subjects (n=83) heard a short sentence, a black circle or square appeared for 200 ms on the central vertical or horizontal axes towards the display margins. They had to press one of two buttons as soon as they had identified the visual stimulus as square or circle, and their response times were registered. The critical auditory stimuli either referred to situations where the figure's trajectory of movement was vertical (e.g., *the ship sinks in the ocean*) or horizontal (e.g., *the car impacts the wall*). They found that response times were shorter when shapes appeared on the axis that was incongruent to the movement axis of the described event, for both vertical and horizontal conditions. Recognition took longer

[29] For further convincing empirical evidence see Slobin (2000) or Papafragou and colleagues (2008).

when the stimuli were presented in the congruent axis, indicating occurrence of a Perky effect. In experiment 2, a new sample (n=81) heard the same stimuli, but had to memorize two cartoon-like images of the mentioned noun-phrase referents, which were drawn in the center of the display concurrently with the unfolding sentence. After six trials in the study phase, they saw an image pair in each of the 12 test trials and had to judge whether they had seen it in the study phase. Richardson and colleagues (2003) assumed a priming effect here, hypothesizing that accurate recognition latency would be shorter when the image pairs were displayed on the axis that is congruent with the event expressed by the main verb. Confirming their hypothesis, subjects did respond faster, indicating that they used spatial information encoded in the verbally described event to build an episodic representation of visual stimuli. Though perceived centrally on the screen, the referent entities were configured to one another along a spatial axis in the online and offline representation constructed by the participants. Both experiments' results support the notion that comprehenders draw on perceptual simulations in the construction of event representations from language. Importantly, the axial configuration of the event participants was induced by the main verb, suggesting that the spatial arrangement came about through dynamics inferred from the expressed action, and not by potentially stereotypical static locations of the NP referents to one another (cf. Estes et al., 2008 below).

Bergen and colleagues (2007) adopted the experiment design used by Richardson (Richardson et al., 2003) and tasked subjects with recognition of geometric shapes at certain screen positions. However, they restricted stimulus space to the vertical dimension. They hypothesized that subjects would exhibit Perky effects after processing single-clause spoken utterances in which vertical translatory motion was lexicalized in the main verb (experiment 1; e.g., *the mule climbed*) or sentences in which the static location of the figure-NP would be above or below the observer (experiment 2; e.g., *the ceiling cracked*). In both experiments, Perky effects were found. Subjects identified stimulus shapes faster when the preceding verbal stimulus expressed motion towards or location on the opposite end of the axis. That is, a shape was recognized faster at the bottom of the screen following an upward stimulus like *the mule climbed* or *the ceiling cracked*.

Estes and colleagues (2008) report commensurable findings. They exposed subjects first to an umbrella term (e.g., *cowboy*) and then a target word (e.g., *hat*), which would be canonically located in a certain position in the vertical axis relative to

the umbrella term. Participants' identifications of single letters in positions congruent with the inferred target location were slower than those where letters appeared in the incongruent position, suggesting that mental representations of static objects are structured analogously to their percept-exemplars.

Matlock (2004), in a series of experiments, set out to test whether subjects comprehended sentences describing fictive motion (Talmy, 2000a) by mentally simulating actual motion. Sentences such as *the road runs along the coast* or *the fence runs across the property* express fictive motion because the finite verb (*runs*) is one of motion, but the figure-NP does not actually move. In her experiments, Matlock (2004) presented short narratives, in which she manipulated characteristics of the conceptual components that are relevant for inferences about the fictive motion cue. For instance, in experiment 1, the context is a narrative about a protagonist's travel on a road through a large desert, which takes multiple hours by car, suggesting to subjects a situation of large-distance travel. Subjects were instructed to decide whether the cue related to that narrative. They took longer to respond to fictive motion cues like *the road runs through the desert* when the narrative context implied large-distance as opposed to short-distance travel. Similar findings hold for experiments 2 (manipulations of figure's implied speed of motion, e.g., a Ferrari vs. Volkswagen bus) and 3 (manipulation of ground terrain, e.g., a rocky dirt road vs. a paved roadway). To be fair, Matlock's (2004) results do not reveal much about processing of fictive motion per se, but rather illustrate how, during comprehension of the context narrative, subjects simulate the relevant components for motion event conceptualization when they construct an event model representation. Subjects' task was focused less on understanding the fictive motion cue as an isolated stimulus, but rather integrating its proposition into the current state of the event model.[30] It is questionable whether Matlock (2004) would have achieved the same results, had the pretext not already specified figure- and ground-based features. Regardless, the case for simulation holds because the response time data was affected in a way that is expectable if one had simulated a figure in motion across a longer or shorter distance[31]. Altogether, Matlock (2004) provides evidence that

[30] Recall that event models are different from working models, in that they already consolidated the previous information of the current episode.
[31] As such, Matlock (2004) presents findings like the mental chronometry reported in Kosslyn, Ball, and Reiser (1978), reviewed in section 2.5.7.

dynamic simulations of motion events incorporate specific features derived from sensorimotor experience and based on Talmy's conceptual components.

The studies discussed so far established evidence that mechanisms of internal information-processing impair attentive processing of external sensory input. When subjects' cognitive system was engaged in simulations of events with pronounced spatiality, their perceptual processing of a briefly flashing visual stimulus was decelerated – a Perky effect occurred. This impedance on stimulus processing is assumed to happen because brain structures for visual processing come to be occupied by simulations: »if a particular part of the retinotopically arranged visual system is being used for one function (…), then it will be significantly less efficient at performing another (…) at the same time« (Bergen et al., 2007: 737), effectively creating bottlenecks of attentive processing in a single modality (Lavie, 2005; Kaschak et al., 2006). However, chronometric measurements related to the detection and recognition of a briefly flashing static stimulus provide, first, only indirect evidence of concurrent – let alone interfering – high-level activity and, second, do not examine visual perception processes as they occur. Before turning to more direct research into visual attention in motion event processing, though, one more elegant study of this niche deserves review.

Meteyard and colleagues (2007) presented their subjects with so-called random dot kinematograms[32] (RDK) while playing them short lists of bare infinitive motion verbs. The RDKs contained coherent upward or downward motion which was either congruent or incongruent with the critical upward or downward motion direction conditions of the verbal stimuli (e.g., sink, lift). Visual controls showed random noise with no coherent motion patterns and control verbs did not lexicalize any motion semantics. Although the detection of coherent motion in a random dot kinematogram depends largely on pre-conscious processing in motion-sensitive areas in the visual cortex, conscious efforts or explicit instruction may improve detection. Meteyard and colleagues' (2007) hypothesized that motion verbs expressing a specific motion direction would bias these automated mechanisms of perception and increase subjects' sensitivity to detect coherent motion signals that congrue with verb meaning.

[32] A random dot kinematogram is a field of randomly positioned dots, some of which move coherently in a specific direction while others move randomly (like white noise, television static). By varying the proportion of coherently moving dots, researchers test one's ability to detect motion patterns amidst noise, exploring preconscious, low-level mechanisms of attention in motion detection.

Subjects were instructed to press one of two buttons if they detected coherent versus random motion in a 150 ms presentation of the RDK. When the orientation axis implied by the motion verb axis was incongruent to the moving dot patterns, participants were less sensitive to this mismatched motion signal in the noise. In other words, mechanisms of visual perception were geared towards not perceiving incongruent motion, suggesting that motion lexeme comprehension biased subjects to not detect mismatched motion. Since subjects' sensitivity to congruent motion was not significantly different from the control condition, however, conclusions about an improvement of detection are invalid. As such, Meteyard and colleagues' (2007) results contradict the screen location-specific interference effects reported in the previous studies. Their subjects did not exhibit weaker perceptual sensitivity when visuo-spatial qualities of internal representations and external stimuli were congruent; instead, perceptual sensitivity in motion detection was reduced when the direction of motion was incongruent with the schematic direction representation activated by the verb — as if the visual system had been expecting something different.

### *Summary of the behavioral evidence*

The previous section demonstrated that people are sensitive to spatial information (like motion direction or trajectory endpoints) when they build conceptual representations of motion events. The reported results suggest that the conceptualization of motion events is supported by perceptual simulations of space.

This conclusion rests mainly on data of response latencies that, as chronometric measures of cognitive processing of external stimuli, were influenced by experimental conditions tied to the conceptual components of motion events. The studies reviewed here agree that certain stimulus conditions, like relative spatial location or implied movement direction, selectively impair attentive processing at specific locations in the visual field (Richardson et al., 2003; Bergen et al., 2007; Estes et al., 2008; Matlock, 2004). In particular, the study by Meteyard and colleagues (2007) suggests that spatial characteristics (like movement direction) are stored in motion event schemata and, even more, that this information is activated by verb lexemes to bias early processes in low-level visual motion perception.

Despite response times offering only a rough and indirect measure of processing, these compelling findings underscore the involvement of the visual system

in processing motion events and highlight the importance of examining direct measures of visual attention, such as eye movements. Consequently, the next section reviews evidence from eye-tracking studies, illustrating attentional sensitivity to spatial information during the perception of visual scenes of motion events, or imagery thereof.

#### 2.6.4.2. Eye-tracking evidence

Eye-tracking research with motion event stimuli is a promising method to examine online processes of perception and cognition, as well as their interactions. The studies reviewed in this section have revealed systematic oculomotor patterns that align with a simulation of movement through space, emphasizing the methodology's potential for studying meaning construction from multimodal input.

Huber and Krist (2004) showed their subjects animated, fronto-parallel visual displays of a ball moving along a surface and then falling off an edge to a specific landing position on the ground. The free-fall trajectory of the ball, which moved at different speeds and fell from different heights, was concealed behind a visual mask. They hypothesized that reasoning about this hidden trajectory, particularly its flight duration, would be improved if subjects used *timing-responsive representations*[33], enabling them to mentally »step through instances of change until the endpoint of the event is reached« (Huber & Krist, 2004: 432). In fact, flight times were estimated more accurately by subjects who exhibited short gaze fixations within the occluded trajectory region of the visual stimulus, as opposed to when they moved gaze straight to the presented endpoint (exp 1-*production condition*). These gaze patterns resembled a dynamic tracking of object movement along a path, suggesting that subjects engaged in visual simulations of the ball's motion to reason about the physics of free fall. Their improved performance underscores the benefits of path simulations for kinematically accurate prediction during visual motion event comprehension[34]. Admittedly, when the visual display remained static, with the ball-figure at initial position, trajectory-tracking eye movements were significantly less frequent than in the dynamic condition, hinting at the possibility that it was the actual movement of the ball that influenced subjects to rely on visual imagery and maintain a dynamic representation for reasoning (see also experiment 5 in Liu, 2009; de Xivry et al., 2008). In experiment 2, however, a different

[33] Timing-responsive representations capture information about the dynamic properties of physical or mechanical systems and are accessed through real or imagined actions that simulate the object's behavior over time (Huber & Krist, 2004).
[34] See also Eisenberg et al. (2018) and Chapter 2.4.3.

sample's accuracy of flight time estimation in animated stimuli was not significantly different between a free- versus fixed-viewing condition, in the latter of which they had to maintain gaze at the position where the ball leaped off the ledge, restricting potential flight path-tracking eye movements. This finding suggests that eye movements are not necessary for accurate estimation of motion features. Still, subjects who executed path-tracking eye movements on their own account in experiment 1 performed better than those who had not and, at the very least, demonstrated reliance on a dynamic representation of trajectory when reasoning about motion.

Kamide and colleagues (2016) recorded their participants' eye movements on visual world displays[35] while exposing them to concurrent spoken sentences about motion events in the scene. The visual world scenes showed a figure- and a ground-entity on either side in frontal view, e.g., an alien called Hillbert and a couch in a room. Exemplified by sentences like *Hillbert will jump* vs. *crawl onto the sofa*, verbal stimuli of the critical condition distinguished an upper- vs. lower motion trajectory to the goal. Although only few participants fixated the path-region in the scenes at all (i.e., the empty part of the image between Hillbert and the sofa), analysis of the vertical coordinates of these fixations indicated a significant difference in gaze allocation as a factor of verbalized trajectory. When the lexical verb implied motion along an upper path, average y-coordinates of fixations were larger (i.e., higher up on the screen) than when verb semantics were associated with a lower path, indicating that sentence comprehension activated an event representation that considered the figure's motion trajectory activated by the verb lexeme, and resulted in congruent, dynamic allocation of visual attention to the external percept.

In a previous study by the same group, Lindsay and colleagues (2013) manipulated manner of motion instead of path. They examined gaze fixations to path regions in visual world stimuli while subjects comprehended sentences implying fast or slow translatory motion of the figure. The stimuli were static images of a figure (e.g., a person) on their way to a goal (e.g., a picnic basket), which was located at the end of a visible path (e.g., a nature trail). Subjects exhibited longer dwell times[36] in the path region when the movement of the figure was described as slow (e.g., *the student will stagger along the trail to the picnic basket*) compared to when the lexical verb denoted

[35] The visual world paradigm is a method in psycholinguistics where participants' eye movements are tracked as they interact with static visual displays of objects or scenes (see Tanenhaus et al., 1995).
[36] »the summed total duration of fixations within a ROI [= region-of-interest, author's note] across a time window« (Lindsay et al., 2013: 4)

fast movement (e.g., *… will run …*). Importantly, summed fixation durations on the figure-entity did not differ between conditions, while looks to the target-object (*picnic basket*) had longer durations when the described motion event implied fast versus slow manner, showing that subjects' eyes reached the goal faster. This additional result is crucial because it casts doubt on one of Talmy's central claims. In Talmy's framework, manner is considered a feature of a figure-object and not an essential component in the conceptualization of motion. Manner is reduced to the status of co-event while path holds conceptual supremacy (cf. *core schema*) (Slobin, 2000: 132). The reported finding shows, however, that manner information does not draw more attention to the figure-object per se but instead to the path- and endpoint-region, suggesting an influence on the understanding of how long a figure occupies a path while moving in a particular way (cf. Papafragou et al., 2008; Matlock, 2004; von Stutterheim et al., 2012). In other words, although the predicate verb encodes manner exclusively, path components seem equally relevant in the comprehension of the expressed motion.

The previous studies illustrate the applicability of eye-tracking in examining the cognitive representations that emerge in online processing of multimodal stimuli in comprehension tasks. It is questionable, however, whether the findings reported so far (Huber & Krist, 2004; Kamide et al., 2016; Lindsay et al., 2013) are in fact indicative of perceptual simulation. Doubts arise because all studies presented visual world stimuli onto which linguistically encoded meaning could be mapped. This became especially apparent in Kamide and colleagues (2016: 808f.), where the introduction of an obstacle into the figure's trajectory (experiment 2) caused a significant increase in path-region fixations compared to stimuli without obstacles (experiment 1). Simply put, subjects looked at the path-region because there was something to look at, better yet, something with respect to which a trajectory could be mapped.

If simulation is conceived as an information-*generating* process, then the tasks administered in the above studies did not require much simulation. The eye-tracking data might just indicate that the bottom-up visual world stimulus and the top-down event schema triggered by language have successfully interfaced in online processing. Simulation, at least the visual type, need not provide (i.e., generate) substantially more perceptual information than can already be extracted from the visual world and upon which linguistic information can be mapped.

On the other hand, longer fixations after slow movement (Lindsay et al., 2013) and higher fixations after upward trajectory (Kamide et al., 2016) reflect that the verbally encoded motion features (something inherently dynamic), as inferred from event schemata, meshed with the perceived visual world (something static) in online comprehension. This inference happens in real-time, as online gaze allocation is influenced immediately. It is not far-fetched to assume that tacit, automatic perceptual simulation, conceived as an information-*activating* process (i.e., activating visuospatial attributes of an event schema), yields these quick inferences (cf. Chapter 2.5.3) and causes immediate differences in gaze durations between conditions. The reported studies remain inconclusive about the precise involvement of simulation, neither do they clarify whether the observed gaze behavior would have been similar, had there not been a visual stimulus to fixate and map onto. The studies reported below tackle this issue.

#### 2.6.4.3. Non-visual gaze in blank screen studies

An early study by Spivey and Geng (2001, Experiment 1) did not show a visual world display onto which event representations could have been mapped. Instead, subjects (n=10) looked at a blank white screen and were tasked with comprehending short narratives. The narratives described situations in which different topic referents[37] were brought into focus at different locations along an implicit axis, inducing subjects to integrate each new proposition of the narrative in relation to the emerging spatial backdrop. For example, in the narrative of the upward condition, »Imagine that you are standing across the street from a 40 story apartment building. At the bottom there is a doorman in blue. On the 10th floor, a woman is hanging her laundry out the window (…)« (Spivey & Geng, 2001: 237), locus of attention moves further up for each successive event. Analysis of the angular orientation of saccadic eye movements revealed that subjects, while gazing at a blank screen, were more likely to execute saccades (> 2° visual angle) in a direction that was congruent with that of internal visual attention relocation through narrative shifts to a new topic referent. This result was robust in all direction conditions compared to the control story (with no implied shifts), i.e., there were more upward-saccades in the upward story than in the control story, as well as between conditions, i.e., there were more upward-saccades in the upward- than the leftward-story.

[37] See Krifka (2008: 247).

Similar findings are reported by Demarais and Cohen (1998). Their participants, like Johnson-Laird (1980), pondered over syllogistic inference problems, such as »a jar of pickles is below a box of tea bags; the jar of pickles is above a can of coffee; where's the can of coffee?« (Demarais & Cohen, 1998: 231). These were presented as auditory stimuli and described common objects from a kitchen context in horizontal or vertical configurations, with the relationships expressed by terms like above and below, or left and right. They assumed that spatial mental model building from verbal input was supported by visual imagery and hypothesized that subjects would exhibit saccades in vertical or horizontal direction at higher rates when the verbal stimulus induced such directionality. Instead of measuring subjects' eye movements on blank screens with a video-based eye-tracking system, they used EOG[38] electrodes to measure muscle impulses from vertical or horizontal eye movements that subjects executed in complete darkness[39]. In two experiments, results showed that horizontal saccade rate was increased during reasoning about left-right problems as opposed to the above-below condition, which, in turn, correlated with a significant increase in frequency of vertical saccades. Interestingly, the more referent objects had to be integrated from the stimulus, the higher saccade rate was measured, although this effect was independent of saccade direction. Additionally, their sample generally executed horizontal saccades more frequently than vertical saccades. Like Spivey and Geng (2001), Demarais and Cohen (1998) conclude that the higher frequencies of saccades in specific directions were driven by automatic internal visualizations of spatial relationships between the stimuli referent objects.

Neither the narratives by Spivey and colleagues (Spivey & Geng, 2001; Spivey, Tyler, Richardson & Young, 2000[40]), nor the syllogisms by Demarais and Cohen (1998) refer to motion events. Effectively, they are sequences of events centered on figure entities localized at different static positions in an emerging scene. What is set in motion is the locus of attention towards the scene. The linear unfolding of the

[38] Electrooculography is a technique for measuring the electrical activity of the eye muscles using electrodes placed around the eyes, typically on the zygomatic and the forehead.
[39] This is to achieve maximal reduction of external visual stimulation. Demarais and Cohen (1998) instructed subjects to keep their eyes open in the dark, while Spivey and colleagues (2000) recorded closed-eye eyeball movements in light.
[40] Spivey, Tyler, Richardson, and Young (2000) previously reported commensurate findings using the same narratives, but a different method. The researchers registered subjects' eye movement directions behind closed eyes. However, closed-eye gaze direction estimation is an imprecise measure, both spatially and temporally, and at risk for researcher bias. Second, mental imagery is more vivid when eyes are closed because external stimulation is largely blocked off (Herff et al., 2022), limiting the applicability of their findings to day-to-day open-eye cognition.

narrative's events invites subjects to create a coherent situation model representation successively and dynamically. Sentence for sentence, the individual events' *figures*, i.e., the topic referents, are configured to one another in accordance with the locative adjuncts in the verbal propositions (cf. Zwaan et al.'s (1995) *event-indexing model*). This is warranted by the above-chance occurrence of saccades in congruent directions, suggesting a scanning of the figure's locations in the internal representation[41]. Most importantly, these systematic eye movements were found even though subjects did not previously see an encodable external visual percept that could have been targeted by fixations or saccades. It is thus likely that they occurred because of perceptual simulations of space in event model representations construed from language. Whether these simulations would constitute generation or activation of information is difficult to assess empirically, since the eye movements were not analyzed as time-locked responses to the verbal input (like in Kamide et al., 2016) but aggregated across the entire trial timespan. Theoretically, since no external visual stimulus was given, the simulation would constitute generation of visuo-spatial representations that support task solving, and, compatible with the visual system, cause eye movements.

***Summary of the eye-tracking evidence***

The eye-tracking studies reviewed here have shown that people move their eyes systematically when they process motion through space. The evidence supports that, first, people simulate motion events dynamically under certain circumstances (i.e., when prompted by animated displays, as in Huber & Krist, 2004), and the component that is simulated for mental representation is path (and manner in Lindsay et al., 2013). Second, these simulations bias eye movements in agreement with idealized path-space in a top-down fashion (Kamide et al., 2016; see also Liu, 2009). Third, the eye movements occur in response to non-verbal video stimuli of motion events (Huber & Krist, 2004; von Stutterheim et al., 2012), as well as to verbal stimuli that, for one, refer to motion events in visual world displays (Kamide et al., 2016; Lindsay et al., 2013) or, for another, describe entities in spatial configurations in the absence of a visual display (Spivey & Geng, 2001; Demarais & Cohen, 1998).

[41] For recent evidence on situation model-triggered eye movements to congruent locations on blank screens, see Johansson, Oren, and Holmqvist (2018).

More generally, these studies highlight the usability of eye-tracking to investigate online processes and representations related to event model building from language, especially when participants are prompted to think about visual space. Within this niche, blank screen eye-tracking has proven a particularly fruitful method to examine mental representation, as eye movements are not steered towards external visual stimulation but spontaneously affected by the contents of the mind. These contents are structured into event models, and, as highlighted by the studies of Spivey and Geng (2001) and Demarais and Cohen (1998), they are built using perceptual simulation (Barsalou, 2008).

### 2.6.5. Summary

People draw on space – a dimension of perception – when they engage in conceptual representation – a dimension of cognition – of motion events. Talmy's extensive research has nurtured a theoretical framework about the conceptual basis of our thinking about motion events. Language assumes a central role in Talmy's framework: it labels spatial concepts that provide schematic structure for our conceptual representations of non-verbally perceived tokens of motion events (Gerwien & von Stutterheim, 2018). Talmy-inspired research has demonstrated repeatedly that conceptual categories, such as figure or ground, substantiate our cognitive apprehension of motion events, whether we perceive them as language, visual stimuli, or through different modalities combined. The empirical research reported further supports that cognition about motion event space is driven by perceptual simulations, which not only affect response times but spontaneous and overt eye movements.

Altogether, motion events have a strong empirical foundation with ample evidence for the psychological reality of their conceptual components, showing that people think about movement in these terms. Motion events are straightforward to use as experimental stimuli, reliably activating event schemata in semantic memory with a discrete set of components. It is through this component-based activation of event schemata that simulation is assumed to initiate, enriching conceptual representations through schema-based inferences. However, the precise manner with which simulation unfolds is difficult to measure directly. Testing a non-invasive and yet feasible approach to such measurement is one of the goals of this study.

## 2.7. Embedding the research question

The comprehension of spoken language is a highly automatized process in which the auditory verbal stimulus is transformed into a conceptual representation. Comprehenders draw on various linguistic levels of description and access their general knowledge bases for decoding, disambiguation, and meaning extraction, ultimately achieving a meaning representation that is largely detached from the surface form of the original input.

Similarly, environmental sound comprehension requires listeners to attend to and analyze the perceptual stream to match source sounds with conceptual knowledge in long-term memory. Through this process, comprehenders form multi-level representations of the input and gradually integrate them into a unified representation, a so-called auditory object. This integration process is guided by schematic knowledge about the source events' typical features and internal structures.

In both spoken language and environmental sound comprehension, stimuli are fleeting, making it necessary to rely on highly automatized and efficient processes for constructing stable meaning representations. This process involves a hierarchical interplay of bottom-up extraction and top-down influences, which supports categorization and predictive inference at multiple representational levels. A key difference between the two modes, however, lies in informativity: in some circumstances, environmental sounds may remain ambiguous or unidentifiable even after extended exposure, while words rapidly activate lexical concepts, significantly narrowing interpretive possibilities of a verbal stimulus. At the same time, environmental sounds automatically evoke more perceptually specific representations, while words cover a broader range of possible perceptual matches.

According to Radvansky and Zacks (2014), the abstract meaning representations generated for language or environmental sound processing are best characterized as event models. These domain-independent, structured representations capture experiential situations and support online processing. Event models arise by matching current sensory input with learned event schemata, and comprehension is established once this bidirectional matching succeeds, indicating that schema-based prediction error is low — a working model has been found.[42]

[42] On a side note, event schemata explain why the two responses to example (1) at the beginning of this chapter are equally felicitous. Regardless of whether Johanna experienced the doorbell-event firsthand or learned about it from her mother, her resultant knowledge state (that someone is waiting at

Radvansky and Zacks (2014) remain imprecise about the details of how the bottom-up and top-down information align in the mind (i.e., prediction error monitoring). They do not specify how the matching of abstract information, which is retrieved top-down, to concrete bottom-up stimuli occurs. One promising proposal that addresses this gap comes from the context of 4E cognition. According to Barsalou's (2008) grounded cognition, conceptual representation occurs through simulations of multimodal, schematic knowledge. Concepts (as knowledge units) and schemata (as networks of related concepts) develop as learned patterns of neural activity, or clusters of co-occurring *perceptual symbols* (Barsalou, 1999; see Ch. 2.5.1). The same brain cells that encode perceptual experience are also in charge of representing experience conceptually. Simulation, in Barsalou's framework, is defined as the activation of perceptual symbols. Since perceptual symbols can be triggered both through bottom-up input analysis or top-down knowledge alike, be it verbal or non-verbal, endogenous or exogenous, the alignment of bottom-up with top-down streams becomes straightforward. The activation of perceptual symbols and the ensuing simulation enable prediction error monitoring by allowing a direct comparison between sensory input and conceptual knowledge. Perceptual simulations of concepts or schemata, including those of events, serve as the representations that underpin comprehension.

In empirical typological research on motion events, the conceptual components and attributes of motion event schemata have been extensively examined, underlining their psychological relevance for cognition and perception (Talmy, 2000a; 2000b). Behavioral studies using response times and eye-tracking data support that people simulate these conceptual components (i.e., figure- and ground-layers) of motion events when integrating them into an event model for meaning representation.

Despite the growing interest and mounting evidence for theories of cognition that emphasize a strong interaction between mental and bodily (or other) processes that extend beyond the cerebral cortex, the picture is far from clear. Disagreement concerns the precise nature of reciprocal influence between verbal and non-verbal representations in conceptual representation. For instance, Speed, Vinson and Vigliocco (2015) criticize the lack of consensus about the »specific mechanisms

the front door) and her subsequent reaction (getting up and walking toward the front door) are best represented by a bell-ringing event model. The underlying event schema provides relevant inferences for Johanna: a person must have rung the bell, expecting a prompt response (doorbell-schema), and waiting at a specific location on the property (house-schema).

responsible for these interactions […] between linguistic and sensory-motor stimuli« (Speed, Vinson & Vigliocco, 2015: 199). Casting doubt on the functional necessity of these reciprocal influences, Hauk (2016) concludes that »the main controversy surrounds the issue about how relevant or essential these contributions [of sensory-motor systems (…) to semantic processing] are, and whether the existing evidence tells us anything interesting about how we represent and process meaning« (Hauk, 2016: 785).

The LASS theory (Barsalou et al., 2008) assumes that language comprehension involves both the activation of concepts through linguistic forms and automatic, concurrent situated simulations. It explicitly takes perceptual simulation as a main computational principle of cognition, arguing that the conceptual representations created for the meanings of verbal utterances are effectively the same as those for interpreting the perceptual experiences referred to in those utterances. It does not matter whether a concept is activated from verbal, sensory, or introspective input, because the activation sets off a situated simulation mechanism tailored to the current task, yielding a conceptual representation of meaning. Similarly envisioning an assembly of multimodal information into a unified mental model, Radvansky and Zacks (2014) hold that utterance meaning is captured by the same cognitive representation that underlies the interpretation of perceptual experience, namely event models. The environmental sound processing literature supports that the levels of mental representation of naturalistic auditory stimuli are compatible with the hierarchical levels assumed in Radvansky and Zacks (2014); such that event segments correspond to auditory events, event models roughly compare to auditory objects, and event schemata share similarities with auditory scenes), while von Stutterheim's (von Stutterheim et al., 2012; Gerwien & von Stutterheim, 2018) research supports the same in language production and comprehension. It follows that comprehension of both verbal and non-verbal auditory motion events involves the construction of unified, domain-independent event models.

Although the evidence presented in Chapter 2.5.7 clearly supports an interaction of verbal and sensory information in conceptual representation, it remains uncertain whether coping with something as symbolic, categorical, and prescriptive as a linguistic system requires situated simulations to become comprehensible or producible. Precisely how do the different modes of knowledge representation (verbal and non-verbal) interact in language comprehension and production, how are they

unified into event models, and under what circumstances may one outweigh the other(s) in cognition?

Accordingly, the following research questions were formulated: Are perceptual simulations constitutive for conceptual representation, whether the input is language or environmental sounds (Experiment 1), as well as whether language is comprehended versus produced (Experiment 2)? Two eye-tracking experiments were conducted to tackle these research questions methodologically, testing whether perceptual simulation is set off automatically to drive conceptual representation of motion events, thus affecting spontaneous oculomotor activity on a blank screen.

# 3. Methodology

Building on Radvansky and Zacks (2014), Talmy (2000b), Gerwien and von Stutterheim (2018), and Barsalou (2008, 2009, 2016), situated simulations in motion event comprehension revolve around the perceptual dimension of a moving figure's direction, which is encoded in event schemata as an idealized path (Talmy, 2000b). Empirical findings and theoretical arguments suggest that simulating this path is fundamental to motion event conceptualization. Accordingly, the present study assumes that processing of motion events relies on visual representations of these abstract, idealized paths. The study's aim is to investigate perceptual simulations in language processing and to determine whether the recording of eye movements can capture simulation as an integral component of conceptual representation.

To examine this, eye movements were recorded of 42 participants during two memory tasks (Experiment 1) and a recognition task (Experiment 2). Participants either understood and memorized non-verbal and verbal events or identified and described non-verbal events. Throughout both experiments, they maintained gaze on a blank screen while an eye-tracker recorded all spontaneous eye movements and other non-visual ocular activity.

***Overview of the chapter***

First, the research questions are operationalized (3.1). Section 3.2 provides a theoretical foundation for investigating simulation through eye movement measurements. This includes an overview of the cognitive (3.2.1) and neurological (3.2.2) bases of eye movements in cognition, followed by a discussion of the blank screen paradigm (3.2.3) as a methodological approach. Section 3.3 then details the empirical investigations, beginning with the experimental setup and technology used (3.3.1), participant recruitment and characteristics (3.3.2), and data collection procedures (3.3.3). The materials and stimuli employed in the study are outlined in 3.3.4. The methodology of Experiment 1 is described in 3.3.5, covering its procedural framework, task design, and experimental conditions, while Experiment 2 is detailed in 3.3.6. Chapter 3.4 summarizes the methodological approach of this study.

## 3.1. Main hypotheses

### *Experiment 1: comprehension*

Are spontaneous eye movements on a blank screen systematically affected by simulations of directionality, which are triggered by motion events in environmental sounds or spoken utterances? The central hypothesis is that during comprehension of both non-verbal and verbal stimuli, eye movements mimic the axial movement direction of the figure of the expressed motion event. To test this, motion events differing in idealized directionality are used as stimuli. The critical condition refers to vertical motion or horizontal motion, and the control condition consists of events without translational motion. If eye movements systematically vary with the experimental conditions, they can be interpreted as indicators of parallel cognitive processes (cf. Ehrlichman & Micic, 2012; Altmann, 2011; Laeng et al., 2014). The prediction is that vertical eye movement is more likely in response to vertical motion events, whereas horizontal eye movements are more likely to follow horizontal motion events, both in contrast with the control condition, where eye movements are not expected to show any systematic patterns. If the predicted eye movements occur in the non-verbal and verbal condition alike, this would suggest visuospatial representation of direction and make a case for perceptual simulation to drive conceptual representation for motion event processing across modalities.

### *Experiment 2: production*

Are spontaneous eye movements on a blank screen similarly affected by simulations of directionality when subjects produce spoken descriptions of motion events? The assumption is that when subjects are tasked to describe motion events, they generate a preverbal message using perceptual simulation and therefore equally exhibit eye movement patterns during language production that vary systematically with the experimental conditions. The hypothesis is that when they identify a vertical or horizontal motion event and prepare to speak about it, subjects are more likely to exhibit vertical or horizontal eye movements as opposed to control events. The motivation for examining both language comprehension and production is the central thesis that simulation constitutes a fundamental mechanism of conceptual representation (Barsalou, 2008): If eye movements reflect the axiality of the motion

events during event comprehension, then these eye movements should be observable during message generation for language production as well.

## 3.2. Eye movements and cognitive processing

Experimental investigations of human cognition face a fundamental limitation: cognitive processes are difficult to measure directly with non-invasive methods (Hauk, 2016). This challenge applies not only to physiological data, such as temporal or spatial measures of brain activity — like electrical potentials (EEG) or blood oxygen levels (fMRI, fNIRS), which are measured with a controlled temporal delay relative to the stimulus — but also to behavioral data, including response times and gaze fixation durations. Chronometric behavioral measures are often used as indicators of cognitive effort, whether in a general sense (e.g., longer response times for condition A compared to condition B suggest differences in mental processing, cf. Shepard & Metzler, 1971) or in a more specific context (e.g., longer fixation durations on a visual stimulus correlate with increased cognitive effort for that element, cf. Just & Carpenter, 1980). However, none of these widely used explanatory variables capture the full complexity of multidimensional cognitive processes occurring in the participant's mind. Instead, they provide approximations along a single dimension, meaning that insights into cognition depend on careful, minimal inferences from directly measurable, associated phenomena. As a result, many findings remain correlative in nature.

The experiments reported here are not exempt from these constraints. How does one examine unconscious perceptual simulations without a magical device that allows extraction of a mind's visual images? Previous research supports a promising method: analyses of participants' oculomotor activity (e.g., pupil dilations, eye movements) and non-visual gaze behavior (e.g., fixation position sequences and relative durations) while they stare at a *blank screen* and solving tasks has been linked successfully to the online, visuo-spatial content on their minds (Fourtassi et al., 2017; Laeng et al., 2014; Richardson et al., 2008; Spivey & Geng, 2001; Weiner & Ehrlichman, 1976; Zikmund, 1973; Antrobus et al., 1964). Chapter 2.6.4.3 discussed evidence that verbally cued visual mental representations trigger systematic eye movements on blank screens. This section deals shortly with both the cognitive assumptions and the neurological basis of how thoughts translate into spontaneous but systematic oculomotor activity.

### 3.2.1. Cognitive basis

In cognition, simulations generate mental representations that have visuospatial features analog to experiential situations – particularly those of motion events. During visual sensation, gaze is directed to new locations in space when shifts of visual attention to such new locations occur (Hoffmann & Subramaniam, 1995; Awh, Armstrong & Moore, 2006; Kowler, 2009). These gaze shifts are usually executed as saccades. Spivey and different colleagues' research (Spivey & Geng, 2001; Spivey, Tyler, Richardson & Young, 2000) demonstrated that mental models emerging from narratives are specific in terms of the spatial relationships between the event participants. Subjects' systematic eye movements in the direction of a novel introduced location relative to that of the anterior entity suggests that they aggregated these entity-tokens into a composite scene representation that retained this spatiality (e.g., Neisser, 1967; Hebb, 1968; Kosslyn, 1994; Liman & Zangemeister, 2012). In other words, the systematic eye movements occurred because, during mental model construction, the mentally represented entities were placed to one another in a spatial relationship that corresponds to the emerging arrangement from narratives (Spivey & Geng, 2001; Spivey, Tyler, Richardson & Young, 2000; Johansson, Holsanova & Holmqvist, 2006; Demarais & Cohen, 1998), single lexemes or lexeme pairs (Estes et al., 2008; Dudschig et al., 2013), or even metaphorical associations (e.g., in basic arithmetic in Hartmann, Mast & Fischer, 2015). In this sense, the eye movements are triggered by shifts of spatial attention to a new location in a mental representation of verbal input, yielding saccades.

### 3.2.2. Neurological basis

The neural mechanisms involved in visual mental imagery run largely on the same neural substrate as visual perception (see Chapter 2.5.2; Kosslyn, Ganis & Thompson, 2001; Dijkstra et al., 2019). Inseparable from vision, oculomotor control resides in a network of brain regions that integrate top-down and bottom-up influences. Here, the *superior colliculus* and the *lateral intraparietal area* are important mediators between peripheral and central activity, acting as a hub for relaying visuospatial attention and oculomotor control. They receive input from visual sensations bottom-up and integrate them with top-down, task-driven demands. This bidirectional accessibility makes it

likely that internally generated images activate oculomotor responses through this pathway (see Johansson, 2013: 37-40). In fact, neurobiological research has shown that the human optic nerve sends impulses from the cortex to the eyes on efferent pathways (Repérant et al, 1989; Shen et al., 2016), indicating that the brain adjusts retinal input with respect to attentional and cognitive demands (Thomas, 1999; Fuchs, 2018: 150; Meteyard et al., 2007). This aligns with predictions of grounded cognition (Barsalou, 2008: 626), in that the »activation of sensory cells during imagery cause this firing of oculomotor cells« (Tillas & Vosgerau, 2016: 467; see also Ishai & Sagi, 1995). Consequently, when external visual stimulation is absent, be it when we close our eyes (Spivey, Tyler, Richardson & Young, 2000; Vredeveldt, Hitch & Baddeley, 2011), dream (Fuchs, 2018: 150; Hobson et al., 2014), or when it is completely dark (Johansson, Holsanova & Holmqvist, 2006; Ehrlichman & Barrett, 1983), this neurological system does not process any external (bottom-up) visual input. It will instead be more responsive to internal stimulation (top-down), e.g., from simulation-related activity of neuron ensembles in the visual cortex, which activated to represent current event schemata. This simulation stimulates neural pathways involved in oculomotor control, thereby generating eye movements that correspond to spatiality of current event representations. Whether these event representations are evoked by verbal descriptions or environmental sounds does not matter (cf. Spivey & Geng, 2001: 240; Barsalou, 2008).

### 3.2.3. The blank screen paradigm

What does matter, however, is that external visual stimulation does not interfere with internally triggered oculomotor commands. We move gaze automatically to visible stimuli, unless we are explicitly instructed or tell ourselves not to. Thus, in investigations of eye movements and cognition, it would be counterproductive to have any bottom-up attractors on the stimulus display at all (cf. Liu, 2009). For this reason, experimental research has successfully used the *blank screen paradigm* (Altmann, 2004) to examine visual representation in online processing (Spivey & Geng, 2001, Exp. 1; Hartmann, Mast & Fischer, 2015; Johansson et al., 2006; Johansson, Oren & Holmqvist, 2018; Brandt & Stark, 1997). In short, eye-tracking with a blank screen paradigm refers to the measurement of eye data on a stimulus display that does not present visual percepts.

The blank screen paradigm is based on the notion that eye movements respond in a systematic fashion to top-down signals from semantic representations. While other studies analyzed gaze shifting to specific locations on blank screens as triggered from long-term memory (a mental map of France, Fourtassi et al., 2017) or working memory (a visual memory trace of a discrete stimulus, Altmann, 2004; Laeng & Teodurescu, 2002; Brandt & Stark, 1997), where fixation positions were interpretable due to a previously encoded visual stimulus, the present study employs the blank screen paradigm in a different manner. Crucially, this study assumes that even in the absence of explicitly encodable stimuli, spontaneous saccades occur in response to visual stimulation — specifically, aligning with the simulated, idealized path directions activated in motion event schemata (cf. Ch. 2.6.4). Consequently, the analysis focuses not on relative eye positions (fixations) but on eye movements (saccades). By employing the blank screen paradigm in this way (cf. Ch. 3.2.3), the study ensures that recorded eye movements are not influenced by previously seen or memorized visual stimuli but instead reflect internal simulations.

## 3.3. The investigations

### 3.3.1. Setup and technology

It is well established that eye movements are closely linked to cognitive activity. However, given the transient nature of cognitive processes and the fact that corresponding oculomotor responses may occur within less than 100 milliseconds, eye-tracking data must be recorded using a system with high temporal and spatial resolution. Such precision ensures the detection of even subtle changes in pupil size or eye position while maintaining stable, continuous recording with minimal data loss. The eye-tracking system employed in these experiments meets these stringent requirements, providing the necessary accuracy and reliability for precise measurement.

In the conducted experiments, the eye-tracking camera (*SR Research Ltd. EyeLink II 1000 Plus*, Host Software v. 5.50) was set to binocular recording mode and a sampling rate of 1000 Hz[43]. The distance between the display screen (24.5 inches,

[43] The *SR Research Ltd. EyeLink 1000 Plus* is a video-based eye-tracking system capable of recording eye data at a sampling rate of up to 2000 Hz, providing real-time gaze data and enabling the detection of even subtle ocular movements (such as microsaccades and glissades) with high temporal and spatial resolution. A specialized camera captures stable images of the participant's eyes and records infrared

resolution: 1920x1080 pixels) and the participants' eyes was 100.5 cm. The chin rest and chair height were adjusted for comfort, and participants were advised to rest their arms comfortably on the table. This was to minimize body movement during recording. The experiment was programmed and executed using *ExperimentBuilder* software package (v. 2.3, SR Research Ltd.) on a high-performance Windows PC.

Auditory stimuli were presented via noise-reducing, over-ear headphones (model *M-Audio HDH-40*) at a comfortable listening volume adjusted individually for each participant. In experimental tasks involving speech production (Experiment 2), participants' voices were recorded using a microphone (model *M-Audio Nova*) positioned directly in front of them, connected via an ASIO audio interface to ensure low-latency transmission.

Participants controlled the experimental progression and provided response times through a controller rigged with five highly time-sensitive buttons (*MilliKey™ Button Box*; keypress latency ~2 ms). Recorded data were exported and prepared for analysis using the *DataViewer* software package (*v. 4.4., SR Research Ltd*).

### *Ensuring eye data quality*

The experiments were conducted in a quiet laboratory room with darkened windows[44]. During recordings, the experimenter remained outside the participant's field of view, continuously monitoring eye-tracking quality on a separate computer. Stationary eye-tracking equipment requires precise technical configurations to minimize signal interference. Signal distortion (technical artefacts or noise) and signal loss primarily stem from suboptimal recording conditions, often caused by participant behaviors or inadequate hardware calibration (SR Research Ltd., 2022). Signal distortion refers to instances in which the eye tracker captures apparent pupil data that do not reliably originate from actual pupil positions, whereas signal loss occurs when no pupil data can be measured, resulting in missing data. A significant problem associated with distorted signals is that they may still display plausible oculomotor dynamics, substantially increasing the likelihood of falsely detected saccades. To prevent such

reflections from the pupil and the cornea (1st Purkinje image). The spatial relationship between the pupil centroid and the corneal reflection is measured at different eye positions on the screen during calibration and incorporated into a mapping algorithm. This algorithm then computes the participant's estimated gaze positions on the screen in pixel coordinates.

[44] Due to scheduling issues, 12 participants in the first session were recorded in a different laboratory with comparable lighting conditions and an identical eye-tracking system.

issues, participants were instructed to sit still, keep their head in the headrest, maintain gaze on the screen, avoid excessive blinking, and refrain from squinting, since even minor movements can invalidate calibration and compromise data quality. If participant characteristics (e.g., glasses or contact lenses) affected recording quality, tracking accuracy was closely monitored, and recalibrations were performed as needed.

### 3.3.2. Participants

The only requirements for participation in this study were binocular vision, native-level German proficiency, and age between 18 and 40 years[45]. A total of 42 subjects participated in both experiments (age: 24±4 years; 11 identified as male (26%), 30 as female (72%), and one as non-binary (~2%). All were right-handed and had normal or corrected-to-normal vision, and the majority was enrolled as students in modern philology programs at Heidelberg University. Although two participants reported bilingual upbringing, they were all German speakers with native proficiency and obtained their secondary education diplomas at German schools. With exception of gender identity, the participants are altogether a balanced and homogenous sample of a student population.

Participants were primarily recruited through brief oral presentations of the experiment in linguistics courses at institutes within the Faculty of Modern Philology at Heidelberg University. Recruitment was specifically targeted at undergraduate and graduate courses whose content indicated clear thematic connections either to experimental methods in linguistics or to cognitive linguistic research questions.

This targeted recruitment attempted to attract participants with an inherent interest in experimental research, ensuring they would be intrinsically motivated[46]. Eye-tracking experiments require a certain physical effort: for extended periods, participants sit upright at a table with their chin resting on a foam cushion and their forehead pressing against a padded metal frame. All the while, they are expected to keep their eyes fixated on a blank screen. Maintaining focus in unnatural posture over a prolonged period is more feasible for participants when they are intrinsically motivated. To bait such participants, various aspects of the study were highlighted using key phrases. The research question was framed as investigating how people

[45] See Dully and colleagues (2018) for a review of age-related cognitive changes.
[46] See Morris and colleagues (2022) for a critical review of intrinsic vs. extrinsic motivation.

infer meaning from auditory input and what cognitive abilities underlie this perceptual process. The study was described as examining auditory memory function, with the task designed as a memory game in which every day environmental sounds would have to be memorized. Compensation was offered in the form of a raffle: among those who completed the study, three gift cards (3 x 30€) were awarded for an online store of their choice.

### 3.3.3. Data collection

Data collection for Experiment 1 was organized into two sessions. All participants attended these sessions approximately four weeks (29 ± 3 days) apart. Session 1 started with the memory task for environmental sounds. Participants were to memorize stimuli in an encoding phase and indicate in a recall phase when they heard novel sounds. After that, the verbalization task was administered (Experiment 2) and participants gave spoken descriptions of the environmental sound events. In Session 2, the memory task for the spoken event descriptions was given (see Table 3-1).

| | | |
|---|---|---|
| Session 1 | Experiment 1 | Memory task with environmental sounds (comprehension—non-verbal condition) |
| | Experiment 2 | Verbalization of environmental sound stimuli (production) |
| | | Self-ratings of visualization intensity |
| Session 2 | Experiment 1 | Memory task with spoken event descriptions (comprehension—verbal condition) |

***Table 3-1**:* Sequence of experimental sessions and tasks.

There were two reasons for this data collection procedure. First, the study aims to compare non-verbal and verbal cognition. To ensure this comparison is valid, it is necessary to design tasks that mobilize cognitive processing that is as unaffected by language as possible (Gerwien & von Stutterheim, 2018; Papafragou et al., 2008). Therefore, task instructions must discourage reliance on linguistic encoding during the environmental sound memory task. If participants were to first covertly verbalize the sounds and then memorize that verbalization, conceptual representation would be primarily verbal. Since this project examines whether event representations constructed in language comprehension are like those constructed for sensory processing, it is essential that task demands do not prompt participants to think in language but instead that they memorize perceptual simulations of the environmental

sound events. For the inferential analytical approach chosen (see Chapter 4), it is necessary to specify how eye movement patterns manifest when participants hold a sensory-based event representation, so that these patterns can serve as a tertium comparationis for those arising from verbally induced event representations.

To meet this requirement, none of the memory task instructions in Session 1 asked participants to use implicit verbalization (e.g., as a mnemonic technique). They were neither required to speak during the task nor to provide language-based responses otherwise (e.g., by pressing a yes/no button). Furthermore, participants were not informed that they would do a verbalization task until they had completed the memory task, thereby preventing strategy building in preparation of the verbalization task. Despite assumptions that non-sense articulation tasks during experiments could suppress subvocalization effects (e.g., by having subjects count, as in Trueswell & Papafragou, 2010), it may be impossible to completely rule out inner speech (Gerwien & von Stutterheim, 2018; 2022).[47] After all, subvocalization during non-linguistic tasks may not influence event processing as much as previously assumed (Papafragou et al., 2008; Trueswell & Papafragou, 2010; Papafragou, 2015: 338) because recent findings indicate no interference with conceptualization in an event description task (Gerwien et al., 2022).

The second reason for using this procedural sequence was to normalize the verbal stimuli for the memory task in Session 2. To ensure that processing of spoken event descriptions would be comparable to that of the environmental sound stimuli, it was necessary to validate first that the sounds were indeed recognizable (see Chapter 3.3.4.2 for details), and second that participants exhibited conceptual agreement in their interpretations. This validation was critical, as a meaningful comparison of cognitively driven eye movements during environmental sound comprehension and language comprehension requires that the stimuli in both modes evoke the same event schemata.

[47] Unless invasive methods like TMS are applied to deactivate the language areas in the brain.

### 3.3.4. Materials and stimuli

All auditory stimuli, both verbal and non-verbal, were processed in *Audacity* (Audacity Team, 2014).

#### 3.3.4.1. Non-verbal stimuli: Environmental sounds

The environmental sound stimuli were either recorded by the author (using a *Tascam DR-05* handheld audio recorder) or downloaded from an online database[48]. All sounds depict common events from domains like household, chores, traffic, animals, nature, human, or sports and recreation.

Each sound was first cut so that the source event immediately set on and then trimmed to a duration as short as possible but as long as necessary for the event to remain identifiable. Where to place these cuts, Marcell and colleagues (2000: 834) have noted, is »clearly more of an artistic than empirical endeavor«. Empirical studies using environmental sounds diverge on this issue[49]. In this study, since recognizability was assessed quantitatively in Experiment 2, cutting and editing mainly intended to minimize variance in stimulus durations for better data comparability. This proved difficult, as some events naturally unfold longer (e.g., a skyrocket), while others express instantaneously (e.g., gunshot), varying strongly in information density.

A second objective of stimulus preparation was to edit the sounds in a way that the to-be-identified auditory object was foregrounded, clearly discernible, and not masked or distorted by concurrent ambience noise (like birdsong in stimuli recorded in nature). This was achieved using frequency-specific filtering and amplification. Unnecessary silence was cut, and noise reduction was applied. All stimuli were exported as Wave files (16-bit signed integer-PCM, Mono, 48kHz sampling rate).

Finally, the exported sound files were normalized to a loudness factor of -33 LUFS[50]. This ensured that all stimuli had the same loudness and that no extreme

[48] freesound.org is a collaborative database of freely downloadable sound samples, including field recordings, musical loops, and sound effects, shared under various Creative Commons licenses (see Appendix A1 for a comprehensive stimulus list and Appendix A2 for *freesound* source URLs and creator IDs).

[49] Those of Röder & Rösler (2003:30) ranged from 400-800ms, Noppeney et al.'s (2007: 600) were on average 800 ± 200 ms long, Marcell et al.'s (2000: 845) items were 2406 ± 1306 ms, and VanPetten & Rheinfelder's (1995: 489) were all cut to 2500 ms. See Dick et al. (2016: 1124) for further discussion.

[50] *Loudness Unit Full Scale* (LUFS) is a standardized measurement that reflects perceived loudness. As opposed to decibel (dBFS), which measures raw signal intensity, LUFS indicates how loud audio sounds, allowing consistent playback volume across stimuli. A smaller negative value indicates greater loudness in LUFS. In some cases, the stimuli were too soft at -33 LUFS and required manual amplification until they were as loud as the other stimuli, based maximum volume peak intensity in dBFS.

changes during perception drew unnecessary attention or even bother participants. For descriptive statistics of the environmental sounds, see Table 3-2.

#### 3.3.4.2. Verbal stimuli: Spoken event descriptions

The verbal stimuli were designed based on the responses of 22 participants (52% of the sample) in Experiment 2. These responses were transcribed and analyzed for recurring linguistic structures. Verbalizations were generally similar in structure and included a subject noun (Talmy's *figure*) and a verb phrase (*activating process*), and a locative adjunct (*ground*). In some cases, motion verbs varied (e.g., *laufen* vs. *gehen*; to walk vs. to go, largely synonymous in German), so the ultimate verb choice depended on preventing overrepresentation of a single verb. Sometimes passive voice was more common than active (e.g., *Reis wird in einen Behälter geschüttet*, rice is being poured into a container). Only when the majority of responses showed passive voice, passive was chosen, in other cases active voice was preferred to highlight the action character of the verbal event. Sometimes the figure-entities were difficult to identify, and corresponding noun phrases varied, while the motion event interpretation was unanimous (e.g., a plastic lid, a pen, a piece of plastic, a ping pong ball, something … fell to the ground). In cases like these, the recorded object was chosen as the figure-NP[51]. Other doubtful cases of variation were resolved in discussions with expert linguists.

Average word count per stimulus was 4, with a standard deviation of 1. The most common syntactic structure that resulted was NP+VP (n=19), NP+VP+PP (n=12), and NP+V+PARTICLE (n=8) or NP+V+NP (n=8). The motion event stimuli expressed various semantic components (cf. Gerwien & von Stutterheim, 2022: 7), mostly Path/object (verb particle) (n=7), e.g., *ein Pferd trabt vorbei* (a horse is trotting past), Goal, e.g., *etwas fällt ins Wasser* (something falls in the water) (n=7), Direction (verb particle), e.g., *eine Münze fällt runter* (a coin is falling down) (n=5), or Location, e.g., *jemand rennt über Schotter* (someone is running on gravel). In other cases, no explicit motion verbs were used, leaving a motion interpretation implicit, e.g., *ein Wasserhahn tropft* (a faucet is dripping) (n=12), although the underlying sound source clearly exhibits translational motion of a figure entity at its core.

[51] If the sound was recorded by the author, the relevant objects were noted in the stimulus file name. If the sound was taken from *freesound.org*, the uploader's sound description was checked.

The verbal stimuli were then recorded by the author with neutral intonation in a noise-insulated recording booth (*M-Audio Nova* condenser microphone, *M-Audio Air* interface routing into *Audacity* at 48kHz sampling rate). All stimuli were read into one large recording, to which low-pass filtering, compression, and equalization was applied, removing bassy frequencies and ensuring consistent intensity. Given that some stimuli had different lengths (utterance word count ranging from 2 to 8), some stimuli were spoken faster and some slower to approximate equal articulation duration. Then, verbal stimuli were cut so that the audio file began exactly with the onset of the first phoneme and ended with offset of the last. In a final step, longer stimuli were sped up to be maximum 2000 ms and shorter stimuli were stretched to be at least 1000 ms. Consequently, all verbal stimuli durations range from 1169-1989 ms, while retaining a naturally sounding rate of speaking. The exported sound files (.wav at 48kHz/16 bit) were normalized to a loudness factor of -35 LUFS. For descriptive statistics of the verbal stimuli, refer to Table 3-2 below.

#### 3.3.4.3. Descriptive statistics of the auditory stimuli

The verbal stimuli have significantly higher loudness levels ($M$[52] = -35.2) than the environmental sounds ($M$ = -33.8, $t$ = -3.6, $p$ < 0.01) but since participants came to both sessions 4 weeks apart, it is very unlikely that such minimal differences in loudness would have introduced confounds. After all, output volume on the headphones was adjusted at the beginning of each session.

| | Environmental sounds | Verbal stimuli |
|---|---|---|
| Avg. duration (± 1 SD) | 2254 (± 644) | 1508 (± 261) |
| Duration range | 740–3913 | 1169–1989 |
| Duration variance ($SD^2$) | 414588 | 68165 |
| LUFS | -34 ± 3 | -35 ± 2 |
| dBFS | -15 ± 5 | -16 ± 1 |

***Table 3-2**:* Summary statistics of auditory stimulus characteristics. Durations are given in milliseconds. LUFS stands for Loudness Unit Full Scale and dBFS stands for decibel Full Scale.

Environmental sound stimuli ($M$ = 2254 ms) also have significantly longer durations than the verbal stimuli ($M$ = 1508 ms; *Two-Sample t-test* $t$ = 8.6, $p$ < 0.01). Within the non-verbal condition (Fig. 3-3, Panel A), horizontal stimuli are longer than verticals (+442ms, $p$ = 0.02), both surpassing non-motion stimuli significantly (verticals by

[52] $M$ denotes the sample mean.

+541ms, and horizontals by +983ms, both $p < 0.001$). The mean durations of environmental sound stimuli are significantly different across conditions ($F(2,61) = 24.5$, $p < 0.01$). In the verbal condition (Fig. 3-3, Panel B), vertical and horizontal stimuli have comparable average durations ($p = 0.94$) but are both significantly longer than non-motion stimuli (verticals by +248 ms, and horizontals by +224 ms, both $p < 0.01$). Again, motion event directionality covaries significantly with verbal stimulus condition ($F(2,61) = 7.2$, $p < 0.01$). None of this is an issue for data analysis per se, as stimulus duration will be controlled for as a possible predictor in mixed-model regression.

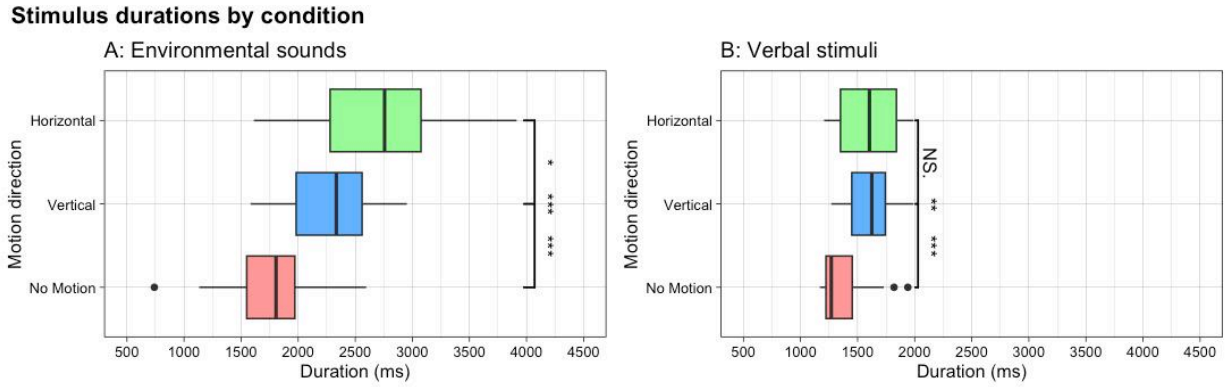


***Figure 3-3**:* Stimulus durations by condition. Stars indicate significance levels (*p < 0.05; **p < 0.01; ***p < 0.001) of duration differences. Environmental sounds (Panel A) show significant duration differences across all conditions, whereas verbal stimuli (Panel B) have comparable durations for the critical conditions but are significantly shorter in the control condition.

#### 3.3.4.4. Blank screen display

The blank screen display was created using the *Paint* program native to Windows 10. It features a centrally positioned square covering an area of 864 px² on the eye-tracking display (see Fig. 3-4). Horizontally, this square occupies 45% of the total display width, while vertically, it spans 80% of the trackable display area. In visual degrees, the grey area covers 13.75 squared degrees of visual angle in the participants' visual field. A square shape was selected to prevent directional bias in eye movements, which could occur if a rectangular shape were used.[53]

The square was rendered in monochrome medium-light gray (RGB 153, 153, 153), with the remaining display area filled in black (RGB 0,0,0). The use of darker colors was intended to reduce eye strain and to minimize the likelihood of squinting, excessive blinking, or participant fatigue due to dry eyes. Additionally, stable display luminance helps to prevent light-induced fluctuations in pupil size, thereby reducing the risk of pupil-related tracking loss.

***Figure 3-4**:* Design of the blank screen display.

[53] The display of the eye-tracking apparatus has 16:9 aspect ratio, thus yielding affordances to execute natural eye movements that are larger in the horizontal than the vertical.

### 3.3.5. Experiment 1

#### 3.3.5.1. Procedure

At the beginning of the first session, each participant provided written informed consent for the research use of their data and was assigned a random three-digit code under which all data were stored anonymously. Information about age and sex was documented, and the dominant eye was determined.[54]

After a brief overview of the procedure (e.g., operating instructions, duration of experiment, two tasks, multiple breaks), participants were informed of basic rules for eye-tracking (e.g., minimizing head movement, maintaining gaze on the screen, avoiding squinting). They then proceeded to the experiment at the eye-tracking computer.

First, headphone volume levels were adjusted. Participants were played a musical chord progression while the experimenter set the output volume to a comfortable level. Then, calibration was performed. It was presented as an attention test, requiring participants to focus on a white dot within a black circle at each new position without moving their heads. Calibration continued until the average estimated gaze estimation error was below 0.5° of visual angle. Following calibration, the memory task was explained on the screen (see Appendix A4), familiarized in a practice block, and then recorded over four rounds (total duration ~15 minutes). Except for the calibration and the recording sequences, participants proceeded through the experiment at their own pace.

The general procedure in the second session (verbal stimuli) was largely identical, except that data collection formalities (e.g., informed consent, personal data) were not repeated.

#### 3.3.5.2. Task and trial design

The data relevant to the main hypothesis of Experiment 1 were collected in a memory tasks. The purpose of the memory task was to activate participants' natural comprehension and encoding processes, as well as their knowledge bases, all of which

[54] Verbatim instructions were, in English translation: »Please form a circle with your thumb and index finger. I want you to hold your hand in front of your face in such a way that it is one upper arm's length away from your face and that your elbow remains at a right angle. Hold your hand in front of your face and look at me through the circle.«

are essential for the recognition and memorization (*encoding*), and the subsequent recognition (*recall*) of the stimuli.

A detailed sketch of the procedure is shown in Figure 3-5. The memory task spanned four blocks. Each block consisted of an *encoding phase* and a *recall phase*, during each of which 12 stimuli were presented in randomized order. In the encoding phase, participants were to memorize and imagine the 12 stimulus events. In the recall phase, most of the stimuli from the encoding phase were repeated, but some were replaced by previously unheard stimuli. The task was to quickly press any button on the controller whenever one of these new stimuli was played. Task compliance thus involved not responding to the stimuli that were remembered. The explicit wording of the instructions was, in English translation:

> »*In the first phase, you will hear 12 different sounds in succession. Your task is to remember these sounds. Each sound suggests an event or action. Imagine these events or actions. […] In the second phase, you will hear another 12 different sounds. Most of them you have already heard and memorized in the first phase, but some will be new. Your task is now to press a button on the controller whenever you hear a new sound*«

This task, which may seem counterintuitive for a memory game[55], is methodologically motivated since it is supposed to prevent limitations and increase the validity of the collected data. The eye data measured here are analyzed as spontaneous, internally triggered saccades that exhibit patterned behavior during stimulus exposure. Requiring a button press upon correct recall could disrupt the occurrence of these internally driven eye movement patterns, because the motor command signaling the moment of event model consolidation (i.e., participant becomes aware of recognition) would require an outward-directed shift of attention towards the button box. Neuroscientific findings have demonstrated interactions between hand and eye movements regarding physiological-motor properties (Engel, Flanders & Soechting, 2002), perceptual aspects (Keetels & Stekelenburg, 2014), or memory performance (Hanning & Deubel, 2018). To avoid that eye movement is coordinated with hand or finger movement to the controller during critical recording periods, a button press was not required for correctly recalled items. This allows multiple measurements of eye movements related to remembered stimuli, during which no efferent, peripheral response was necessary.

[55] The correct response requires a no-go, that is, an inhibition of response behavior.

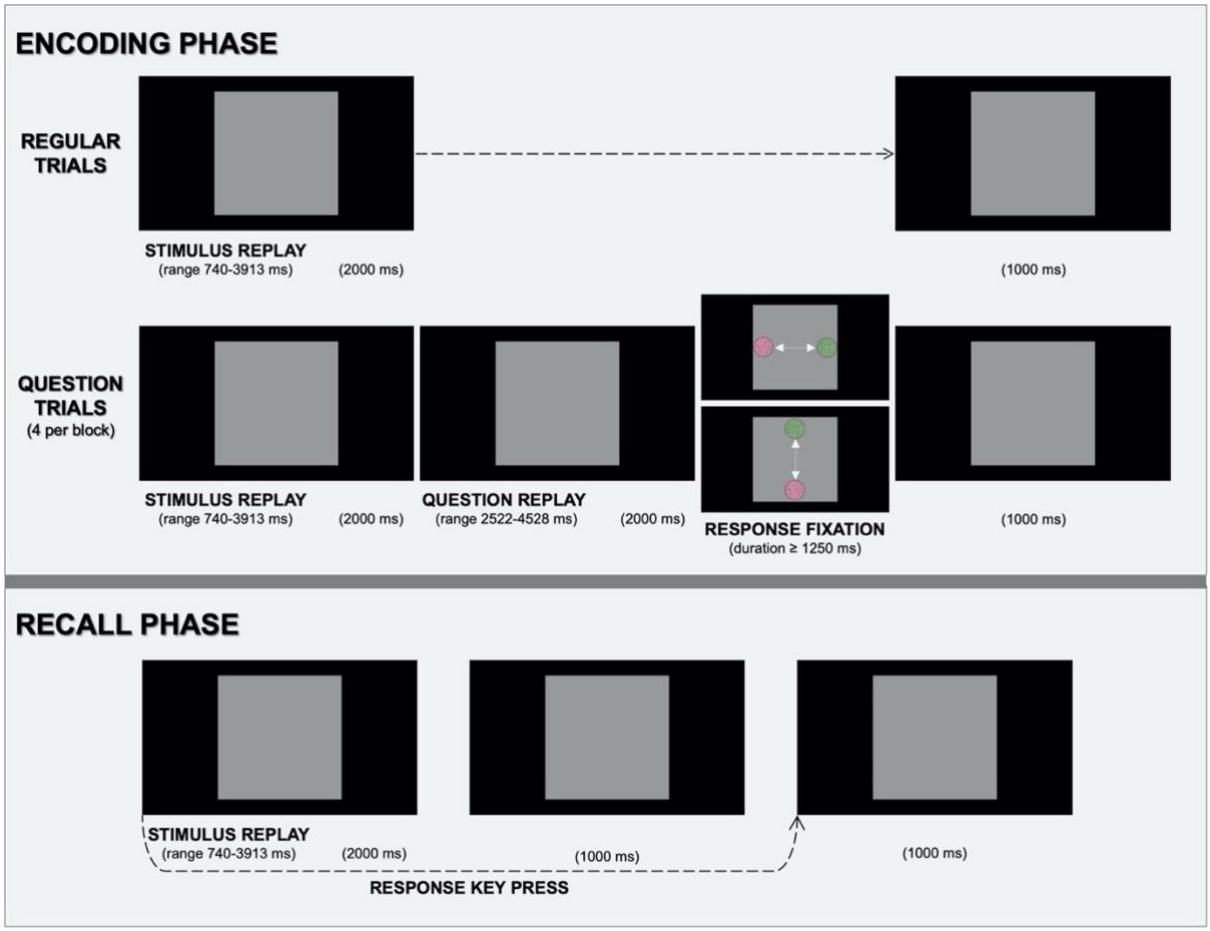


***Figure 3-5**:* Trial structure of the task in Experiment 1. Panels that are not labeled but contain duration information are silent displays of the blank screen. Time proceeds from left to right, and data for analysis was recorded in the trial periods depicted in the left-most panels 'Stimulus replay' and '(2000 ms)'. The first row shows the procedure in 'regular trials' of the encoding phase, where sounds were played in succession. The second row shows 'question trials', in which sensory-based questions were asked shortly after stimulus offset and gaze responses were demanded. The third row illustrates the trial procedure in the recall phase.

The relevant data for analysis are recorded during the stimulus replay phase and the subsequent 2000 ms silent segment (see the left-most panels in Fig. 3-5). Theoretically, this combined epoch represents immediate perceptual and cognitive processing of the stimulus, in both encoding and recall phases. The methodological rationale for analyzing these data is that epoch durations remain consistent in for each stimulus in both phases, yielding comparable exposure times — unless the trial is prematurely aborted in the recall phase by a keypress.

The 2000 ms segment of silence after stimulus presentation was intended to allow for additional, undistracted conceptual processing and memory encoding. Presumably, these processes would be finished by around 1000 ms post-stimulus onset (cf. Kutas et al., 2007; Hagoort et al., 2004; Indefrey, 2011). Furthermore, certain environmental sound stimuli might be identified only after they have finished playing,

so these additional 2000 ms captured eye movements tied to critical cognitive processes.[56] Finally, an extra 1000 ms of silence at the end of each trial allowed participants to disengage from the previous stimulus without feeling hurried into the next trial.

### *Sensory questions*

In the encoding phase, participants sometimes responded to sensory-based questions (see Appendix A3) (Gerwien et al., 2024; Laeng et al., 2014; Noppeney et al., 2007). The explicit instruction was:

> »From time to time, you will be asked questions about the events. You answer 'yes' by fixating your gaze on the green dot, and 'no' by fixating your gaze on the red dot. Please answer intuitively and do not spend too much time thinking about it.«

These yes/no questions were presented on a third of all encoding trials (4 out of 12 per block) and played as audio recordings after a post-stimulus silence of 2000 ms (see Fig. 3-5). The questions targeted different conceptual properties related to the event schema of the preceding stimulus (see Table 3-6).

| Category | Question (English translation) | Duration |
|---|---|---|
| Temporal extension | Does this event usually span a relatively long period of time? | 4475 |
| Typical location | Would one typically observe this event outside a building? | 4528 |
| | Would one typically observe this event in a building? | 4034 |
| Typical participant | Is a human being involved in this event? | 2522 |
| | Is an animal involved in this event? | |
| Instrument use | Is this an event in which a person uses some sort of tool? | 3950 |

***Table 3-6**:* Structure of the sensory questions. For German originals, see Appendix A3.

Two colored dots, each subtending ~3.2 degrees of visual angle, appeared on the blank screen after the question (see second row in Fig. 3-5) and participants responded by fixating their gaze on the one with green visual noise (for 'yes') or red (for 'no'). They

[56] In the recall phase, an additional 1000 ms segment was included (middle panel) to provide participants with extra time to determine whether they had previously heard the sound. Memorizing by sound is not a routine task, and a 2000 ms decision window may not be long enough. Without this additional time, there was a risk that novel sounds would be mistakenly classified as 'heard before' simply because participants failed to respond within the given timeframe. Extending the response window therefore aimed to reduce accidental type I errors (false positives).

were positioned opposite each other on either the vertical or horizontal central axis and were counterbalanced in these positions, rendering their placement unpredictable. Participants saw either of the four positional permutations just once per block. Fixation duration had to be at least 1250 ms for it to be registered as the volitional response.

The sensory questions served multiple purposes. Primarily, they were attention checks, ensuring that participants remained focused on processing all stimuli attentively. Focused processing of the stimuli was essential for interpreting eye movement data in terms of cognitive processes. If participants' attention was diverted by task-unrelated thoughts (Steindorf & Rummel, 2020), the validity of such inferences would be questionable. The distribution of questions across stimuli was pseudorandomized, as not all question types were applicable to every stimulus. The assignment of questions to stimulus conditions was counterbalanced, making it unpredictable from stimulus type. Consequently, the questions always surprised participants, fulfilling their purpose as attention checks.

The second motivation to use these questions was to induce activation of sensory representations for conceptual processing of the stimulus events (Gerwien et al., 2024). This was to mitigate the risk that, due to superficial processing or verbal labeling, a simulation might not capture sufficient attention to trigger spontaneous oculomotor activity (Barsalou et al., 2008). In other words, the questions aimed to bias participants toward perceptual simulations of events, such that it becomes the representational medium for both response selection and stimulus memorization.

Note that the eye data from the sensory questions and the response were not analyzed, mainly because the question was played as a voice recording (setting off automated language comprehension processes) and the response, though a sensory-based decision, required a quasi-linguistic yes/no-answer (i.e., symbolic relation of green dot = yes and red dot = no). It seemed futile to analyze the eye data in these trial epochs because both non-verbal and verbal cognition coalesce in the response, and eye movements may be affected by coincidental language processing.

On a final note, response accuracy was not important for the analysis. Consequently, no error feedback was given, neither on the button presses in the recall phase, nor on the responses to the sensory questions.

#### 3.3.5.3. Experimental conditions and independent variables

In the memory task, data recorded in two consecutive trial segments captured eye movements during immediate stimulus processing. Eye movements from these segments will be analyzed for systematic directional patterns along the vertical or horizontal axis. The primary objective of the analysis will be to determine whether the typical movement direction of the figure in each stimulus influenced the occurrence of these directed eye movements during immediate processing. Consequently, movement direction served as the main experimental condition.

##### 3.3.5.3.1. Movement direction: vertical vs. horizontal vs. non-motion

The 64 stimuli are divided into three groups based on stimulus conditions[57]. Stimuli in the critical conditions depict a motion event in which an entity moves along the vertical axis (n=16, e.g., a skyrocket during fireworks) or events with horizontal movement (n=24, e.g., a passing motorbike). The control condition (n=24) includes stimuli that cannot be interpreted as translational motion events, let alone as directed locomotion (e.g., a meowing cat). Note that the horizontal movement and control condition contain more stimuli than the vertical condition. This is because some novel, non-repeated stimuli were required in the recall phase, so that participants could justifiably press buttons and comply with the memory task. Table 3-7 (Ch. 3.3.5.4) depicts how stimuli were distributed among conditions in all four blocks.

For reasons why motion events make convenient stimuli in this experiment, refer to Chapter 2.6.2. With regard to the specific comparison of vertical and horizontal movement direction vs. non-motion stimuli, the reviewed research supports that subjects are sensitive to these axial distinctions, both in low-level visual perception (Meteyard et al., 2007; Estes et al., 2008; Spivey & Geng, 2001; Liu, 2009) and higher-level processes like event unit formation (Gerwien & von Stutterheim, 2018) or memory encoding (von Stutterheim et al., 2012). Also, sensing up- or downward and horizontal motion directionality is a fundamental environmental experience that can be readily encoded in German, where direction is typically expressed as a separable adjunct to the verb (a *satellite*; Talmy, 2000a).

Importantly, when it comes to hearing motion, humans do not perceive vertical motion as effectively as horizontal motion (McDermott, 2013: 150f.). This is partly due to the shape of our heads and ears, whose lateral placement makes us physiologically

[57] For a full list of the stimuli, see Appendix A1.

more adept at sensing horizontal movement than vertical movement. In the most extreme case, this could cause a species-specific bias to interpret heard movements primarily as occurring in the horizontal plane. Because we are less sensitive to vertical motion acoustically, any vertical eye movement patterns in the vertical condition are more likely to result from a visuospatially enriched conceptual representation of movement direction rather than from perceiving the relative vertical location of an external sound source — especially since participants wear headphones delivering sound laterally. Moreover, all environmental sounds were played in mono, effectively ruling out the possibility of horizontal gaze shifts triggered by a unilaterally louder ear signal (e.g., Doppler effect).

#### 3.3.5.3.2. Stimulus modality: verbal vs. non-verbal

All stimuli were presented as non-verbal environmental sounds and spoken event descriptions. Comparing the comprehension of motion events in the two modalities directly targets the research question of whether perceptual simulation is employed as a general mechanism of conceptual representation that cuts across different modalities. In order to be able to contrast verbal vs. non-verbal modalities in the predicted blank screen eye movements, visual or written text stimuli were not an option, because they would elicit eye movements in form of visual gaze (Altmann, 2004; Laeng & Teodorescu, 2002). Motion event representations had to be evoked without introducing visual stimuli like videos or written text. Besides vision or language, no other representational modality is suited as well as audition to reliably activate motion event representations in controlled laboratory settings[58].

#### 3.3.5.3.3. Self-assessed visualization intensity

Individual variation is expected in how participants approach the tasks in this experiment because they may exhibit different cognitive styles (Riding, 1997; Kozhevnikov, 2007). For instance, participants may differ to what extent simulations capture their attention when they conceptually represent events during comprehension or production. This idea of distinct processing modes is not new (Paivio, 1971) and many studies have quantified individual differences in, e.g., mental imagery ability (Sheehan, 1967; Marks, 1973; 1995; Andrade et al., 2014), allowing for assessment of

[58] Except perhaps the haptic sense or proprioception, although inducing motion event representations like that would be highly challenging in an eye-tracking paradigm, let alone regarding ethical concerns.

visual imagery strength during cognitive tasks. Many of these questionnaires consist of self-ratings and their reliability remains debated (see, e.g., McKelvie, 1995; Schwarzkopf, 2024). Nevertheless, they have consistently revealed considerable variation in individuals' reports about how vividly they imagine concepts (Marks, 1973; McKelvie & Demers, 1979; Charlot et al., 1992; Lovell & Collins, 2002; Keogh et al., 2020). Consequently, there is no consensus on how mental images are phenomenologically experienced in the general population. Some experience them as if they were 'seeing' them in overlap with external visual percepts, while others locate them off their visual field 'behind their eyes', like a barely noticeable imagistic sketch inside the mind (Schwarzkopf, 2024). However, a »complete overlap between imagery and perception is not to be expected given that it is crucial for our cognitive system to be able to distinguish the two« (Johansson, 2013: 28). Thus, people possibly vary along a gradient in how they become aware of these internally generated representations and, consequently, how strongly their attention may be drawn perceptual simulations, and this may also reflect in the self-reports of the present sample. None of the eye-tracking studies reported in Chapter 2.6.4 controlled for this potential heterogeneity in the experience of mental images.

However, measuring conscious mental imagery — i.e., the explicit experience of internal images before the mind's eye — and unconscious simulation — i.e., activation of perceptual symbols across cortical networks for concept representation — are two different but closely related notions. First, aphantasics, despite being unable to voluntarily generate mental images, still report experiencing involuntary visual imagery in flashes or dreams, that is, unconscious simulations breaking through to consciousness (Zeman et al., 2015). This suggests that even in those unable to perform imagery voluntarily, perceptual simulations still seem to run unnoticed and possibly support conceptual representation without conscious awareness.

In line with this distinction, Barsalou (2008; 2016) has argued that simulations often remain defocused in conceptual processing and only influence cognition without producing a consciously attended mental image (cf. Nanay, 2021). At the same time, empirical evidence demonstrates that visual mental imagery is caused by activity of largely the same neural systems that are involved in visual perception (Dijkstra, Bosch & van Gerven, 2019; although see Bartolomeo, 2008), including those responsible for perceptual simulations (i.e., activation of perceptual symbols) in the visual cortex (Andrade et al., 2014). Recent research converges on the notion that there is a link

between the intensity of mental imagery and visual cortex excitability (Keogh, Bergmann & Pearson, 2020; Dance et al., 2021; Charlot et al., 1992). If the visual cortex is more excitable, it is more responsive to internal stimulation, and the resultant mental imagery will be perceived as more vivid. According to a proposal by Kvamme and colleagues (2024), »individuals with a predisposition for vivid mental imagery might naturally gravitate toward focusing inward, as these rich inner experiences could capture their attention and foster deeper engagement« (Kvamme et al., 2024: 3). Thus, variations in cortical excitability may explain individual differences in how vividly people experience their mental images, how easily their attention is captured by perceptual simulation, and how readily they rely on perceptual simulation for conceptual representation in cognitive tasks.

Given the documented individual variation in imagery vividness, it is reasonable to assume that participants in this sample also exhibit such variation. Some individuals, due to habitual preferences or a predisposition for a particular mode of conceptual processing (e.g., language and situated simulation, Barsalou et al., 2008; visual vs. verbal, Paivio, 1971; see Koć-Januchta et al., 2017; Toomey & Heo, 2019 for evidence for such preferences) or specialized expertise (Blazhenkova & Kozhevnikov, 2010) may favor one processing mode over another during the experimental tasks. Their disposition to draw on perceptual simulations may vary, and it would be methodologically and analytically unsound to consider the sample homogenous in this respect. Ultimately, the stronger the habitual reliance on simulation, the more oculomotor activity might occur — likely due to increased cortical activity in visual system (Dance et al., 2021) and more attention being captured by emerging visual representations. The involuntary capture of internal attention by perceptual simulations may therefore be associated with the eye movement patterns hypothesized in this study (see Chapter 3.2; Gurtner, Hartmann & Mast, 2021: 10).

To account for this variability, participants' visualization intensity was measured using self-reported ratings. At the end of Session 1 (after Experiment 2), participants were asked to rate the vividness of their mental imagery separately for Experiments 1 and 2 on a 5-point scale[59]. Higher self-reported imagery scores likely indicate greater visual cortex activation during mental representation, suggesting a stronger inclination

[59] The verbatim question was: »How vivid were your visualizations of the heard events, on a scale from one ('not very vivid, weak') to five ('very vivid, strong')?«.

towards perceptual simulation in conceptual processing. These ratings were therefore included as an independent variable in the statistical models (see Ch. 4).

#### 3.3.5.3.4. Task phase: Encoding vs. recall

Designing the memory task with an encoding and recall phase was primarily motivated by the need for repeated measures, thereby increasing the amount of data available for statistical analysis. The empirical motivation was to induce participants to activate the same event models in both phases, allowing comparison of eye movement data from both encounters. However, are both encounters with the stimulus identical, or do they differ in terms of cognitive processing?

Assuming they successfully memorized a stimulus, participants' knowledge dispositions for event model construction differ in recall and encoding. During encoding, they must establish a working model and encode the stable event model into memory (first encounter). When the stimulus is presented again during recall, recognition relies both on bottom-up processing of the auditory input and on the memory trace of the previously stored event model — the one that captured the meaning of the sound on first encounter. In other words, the first encounter involves perceptual processing for working model creation and encoding (strongly bottom-up driven and with top-down activation of event schema), while the second encounter involves perception for recognition (minimal bottom-up processing, quickly transitioning to memory retrieval of a previously established event model). According to grounded cognition, situated simulations for the conceptual representation of a stimulus event during encoding should be reactivated in memory recall (Barsalou, 2008: 626).

One potential caveat is that recognition in the second encounter might rely on echoic short-term memory for superficial acoustic features (e.g., a loud bang) or linguistic surface forms (e.g., remembering a specific word) rather than on deeper event model construction. The sensory questions counteracted such superficial strategies by requiring participants to be alert for questions, thereby increasing the likelihood of preparatory event schema activation. In fact, Röder and Rösler (2003) demonstrated that semantic encoding strategies, as opposed to encoding based on acoustic properties of environmental sounds, led to superior recognition performance across all their participant groups (sighted, congenitally blind, and late blind). However, their memory task included a significantly larger stimulus set, with 59 items in the study

phase and twice that amount (59 old + 59 new) in the recognition phase. In their study, memorizing the environmental sounds by clustering them into larger conceptual structures (e.g., sounds associated with traffic) may have proven more efficient because it reduced cognitive load. Nonetheless, Röder and Rösler (2003) provide evidence that recall performance improves when semantic representations, such as event models, and not superficial perceptual features are encoded.

#### 3.3.5.4. Counterbalancing and pseudorandomization

Each participant was presented 64 unique stimuli in total, with 12 per encoding and recall phase. All stimuli of the vertical condition (n=16) were repeated in the recall phase. For the 16 horizontal and 16 non-motion stimuli, half (8 per condition) were repeated in the recall phases, while the other half were replaced with novel stimuli.

Each encoding and recall phase contained 4 stimuli from each condition. In the four recall phases, either three or five stimuli from the horizontal or control conditions were replaced with new ones in alternating phases (see Table 3-7).

| Condition | | Critical-vertical | | Critical-horizontal | | | Control-no motion | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task phase | | Encoding | Recall | Encoding | Recall | | Encoding | Recall | | Encoding | Recall | |
| | | | repeated | | repeated | novel | | repeated | novel | | repeated | novel |
| Block | 1 | 4 | 4 | 4 | 3 | 1 | 4 | 2 | 2 | 12 | 9 | 3 |
| | 2 | 4 | 4 | 4 | 2 | 2 | 4 | 1 | 3 | 12 | 7 | 5 |
| | 3 | 4 | 4 | 4 | 1 | 3 | 4 | 2 | 2 | 12 | 7 | 5 |
| | 4 | 4 | 4 | 4 | 2 | 2 | 4 | 3 | 1 | 12 | 9 | 3 |
| Total | | 16 | 16 | 16 | 8 | 8 | 16 | 8 | 8 | 48 | 48 | 16 |

***Table 3-7**:* Counterbalanced distribution of stimuli in different conditions across blocks and task phases.

All stimuli of the vertical condition were repeated to obtain more measurements. In contrast, not all horizontal or control stimuli were repeated.[60] Nevertheless, participants were exposed to an equal number of trials with equal numbers of stimuli from each condition. Stimuli of the vertical condition were always identical to those in the encoding phase, meaning participants should not have pressed any keys, whereas the horizontal and control conditions featured either 3 or 5 genuinely different events (i.e., novel recall-stimuli).

[60] The ones chosen to be repeated had been judged as easily recognizable in an unreported pilot study (n=21), whereas the novel ones, which were only played in recall, were more difficult to recognize.

This variation in block design primarily aimed to enhance the authenticity of the task as a memory game. A memory game would be less plausible if each round contained only a small, fixed number of repeated trials. Introducing variation made the experiment more dynamic and engaging, with the element of surprise reducing the likelihood that participants develop suspicions about a potential research objective. At the same time, greater engagement in the task enhances intrinsic motivation and attentive listening, leading participants to think as though they were actively playing a game rather than complying with instructions of a laboratory experiment. Overall, this gamification not only masked the research objective but also incited natural processing mechanisms yielding spontaneous and ecologically valid data.

The distribution of stimuli across blocks was carefully controlled through pseudorandomization, avoiding systematic co-occurrences across encoding and recall phases. Several criteria guided this process.

First, the stimuli in the same block had to be clearly distinguishable. Some environmental sounds — particularly those depicting horizontal motion events — shared similar occurrence domains (e.g., a passing car vs. a passing motorcycle), or acoustic properties like pitch (e.g., ambulance sirens vs. fire truck sirens), or timbre (e.g., marching vs. jogging, or walking on a wooden staircase vs. a wooden floor). Because of these similarities, confusion was more likely if they appeared in the same block. Additionally, certain stimuli could be construed as related (e.g., running faucet vs. pouring a glass of water), and according to Röder and Rösler (2003: 29), cue items that are semantically linked to memorized items increase the risk of false memories. Additionally, the novel stimuli introduced in the recall phase were selected so that both acoustic and semantic overlap with those from the encoding phase was avoided.

Complicating matters further, the distribution of sensory questions across encoding phases needed to be balanced. This ensured that each question per block targeted distinguishable conceptual attributes and was equally distributed across stimulus conditions (e.g., to avoid that two instrument-questions occur on horizontal stimuli in encoding phase of block 2) throughout the whole task. Also, to prevent participants from becoming discouraged, sensory questions were limited to easily identifiable sounds.

### 3.3.6. Experiment 2

#### 3.3.6.1. Procedure

In Experiment 2, speech and eye movement data were recorded while participants listened to the environmental sound stimuli again and verbalized the interpreted events. Because Experiment 2 followed the non-verbal memory task of Experiment 1 in Session 1, participants were given a break of a few minutes during which the microphone was set up. Participants then returned to the eye-tracking computer to read the instructions of the verbalization task (see Appendix A4). If there were no outstanding questions about the task instructions, calibration was repeated, and recording commenced without any practice trials since subjects had already become accustomed to the general procedure. After the verbalization task, the eye-tracking recording was terminated.

#### 3.3.6.2. Task and trial design

Verbatim task instructions were as follows:

> »You will hear a sound. As soon as you can describe it, press a button on the controller to start the voice recording. Speak clearly and distinctly, but spontaneously and naturally into the microphone. The recording will end automatically, and the next sound will be played.« (see Appendix A4 for German original)

The task was divided into four blocks. In each block, 16 stimuli were played in randomized order. The same blank screen as in Experiment 1 was shown and participants were again instructed to maintain gaze within the grey square. Stimulus materials are described in Chapter 3.3.4 and the same stimulus conditions apply (horizontal, vertical, control). Conditions were counterbalanced across blocks (4 verticals and 6 each of horizontals and distractors). No sensory questions were asked, and no error feedback was given on the verbalizations.

A sketch of the trial structure is depicted in Fig. 3-8. Contrary to the instructions, voice recording started automatically at trial onset and before stimulus replay, recording all verbal responses during a trial. Response key presses interrupted audio replay, allowing participants to respond without waiting for the stimulus to finish. They were instructed to verbalize the essential event, without emphasizing too many details. As soon as they began speaking, the system detected their voice, and, unbeknownst

to the participants, an automated timer set off to limit their response to 6 seconds. The end of the voice recording was signaled with a success-sound icon. If participants did not press a key within 4 seconds from stimulus offset, they probably did not recognize the stimulus on first encounter and were given the chance to repeat it once. If they failed to recognize it after repetition, they were to press, say 'no idea', and proceed to the next trial. Neither repeated, nor unrecognized trials were considered for analysis.

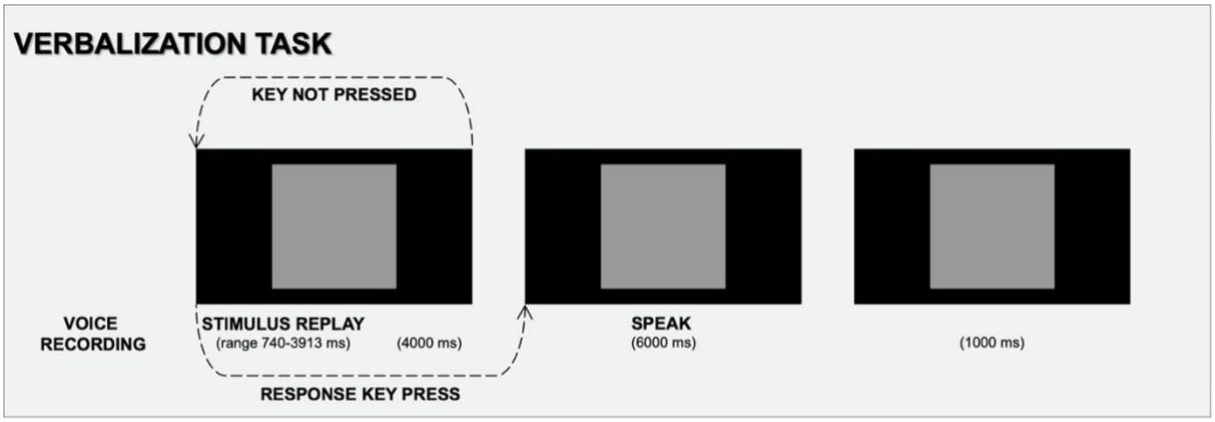


***Figure 3-8**:* Trial structure of experiment 2. Gaze data was recorded throughout the whole trial. Keypresses fast-forwarded the procedure to voice recording. Absence of keypress after 4000 ms post-stimulus offset repeated the trial once. Participants had 6000 ms to speak their response.

The verbalization task serves two distinct purposes in this study. The first concerns the analyzability of the eye movement data. The primary goal of Experiment 2 is to examine whether eye movement patterns observed during conceptual representation for comprehension in Experiment 1 would also appear during conceptualization in language production. In other words, it attempts to find evidence in eye movements for the hypothesis that the mental representation generated during a comprehension process is like the one constructed during message generation.

Participants begin each stimulus verbalization by constructing a conceptual representation of the environmental sound (i.e., the preverbal message) and transform it into an articulable utterance in a multi-step process (Levelt, 1989; von Stutterheim & Nüse, 2003). The trial procedure described here (see Fig. 3-8) allows for recording of eye movements associated with these stages of speech production. From the onset of the auditory stimulus to the button press, participants mainly perceive the stimulus and construct the preverbal message (later termed the *audio replay* phase). The button press indicates readiness to speak, signaling that the first increment of the verbalizable message has been formulated. The time interval between keypress and voice onset (*pre-voice* phase) corresponds roughly to the formulation stage in language

production, including lemma retrieval and syntactic encoding for further increments. Speech onset marks the point at which the first utterance increment was fully phonologically encoded and articulable (*articulation* phase). Empirical research on language production has used button response times and speech onsets as key reference points for segmenting trials into cognitively relevant processing stages (Schriefers et al., 1990).

Importantly, these stages are not viewed as strictly isolated, sequential steps, where each corresponds to a discrete cognitive operation (cf. Ch. 2.1.1). Instead, they represent milestones in the individual's trial-specific subprocesses of online speech production. The timestamps of the behavioral responses during the verbalization task indicate that participants sampled and processed sufficient information to proceed to the next global step of task execution. In the present design, the keypress signals that conceptualization is completed to a degree that allowed the first conceptual component to be passed on to the formulator, and what falls between keypress and voice onset are cognitive processes largely dedicated to linguistic encoding. The eye data from these distinct epochs will be interpreted as indicative of cognitive activity dedicated to these subprocesses (cf. Ch. 5).

Beyond these methodological and analytical considerations, the verbalizations served to evaluate response accuracy and stimulus recognizability, as well as perform internal validation of the verbal stimuli for Experiment 1. First, participants' verbalizations reveal their interpretation of the stimuli. This allows for the assessment of whether they understood the sounds as the intended events, had only a partial understanding, or failed to recognize it altogether. Items that showed inconsistent interpretations (e.g., low naming agreement) or were frequently unrecognized (i.e., high number of 'no idea' responses) were later excluded from analysis, either by subject or across both experiments.[61] Second, these verbalizations were used to validate and normalize the linguistic stimuli for the verbal memory task of Experiment 1, as detailed in Chapter 3.3.4.2. Finally, analyses of participants' utterances enable inferences about event construal and perspectivization (cf. Ch. 4.5.7.3). This reveals which event components were perceived as sufficiently salient to be verbalized, such as path information, potential agents involved in the action, spatial context, or possible consequences of the event.

[61] After all, if a participant did not identify a sound during verbalization, they likely did not recognize it during the memory task either. An analysis of eye movement data from these specific trials would not make sense in either scenario.

### 3.3.7. Summary of the study design

The present study employs an experimental design in which the outcome variable is systematically manipulated by independent variables with distinct category levels (e.g., verbal vs. non-verbal; vertical vs. horizontal vs. non-motion).

Experiment 1 follows a 2x3 factorial design, with *stimulus modality* as the first factor (two levels: verbal vs. non-verbal) and *movement direction* as the second factor (three levels: vertical, horizontal, and non-motion). Each stimulus is thus classified based on both modality and motion event characteristics, resulting in six experimental conditions. Participants' self-reported *visualization intensity* is assumed to be associated with their oculomotor activity. While an additional factor *task phase* (encoding vs. recall) is not expected to systematically influence the outcome variables, its effects were explored post-hoc.

Similarly, Experiment 2 assumes that eye movements are differentially influenced by the movement direction implied in participants' event descriptions, but with a specific focus on language production. Although no specific hypotheses are formulated regarding the effects of verbal responses on eye movements, potential influences of construal, as reflected by participants' verbalizations, will be examined in a post-hoc analysis.

## 3.4. Summary

Chapter 3 outlined the study's hypotheses, followed by a rationale for using blank screen eye-tracking to investigate mental representations in cognitive processing. In support of this method, relevant cognitive and neurological foundations that underscore how subtle eye movements can reflect underlying mental processes were briefly discussed.

Next, a detailed account of the experimental setup and recording procedure was given, including the technologies employed and measures implemented to ensure high data quality. The chapter also described the participant sample, recruitment strategies, and the plan behind the particular data collection procedure.

The core of the chapter is focused on two experiments. Experiment 1 probes comprehension of motion events across stimulus modalities and is disguised as a memory task. Its procedure, task design, experimental conditions, and measures taken to ensure validity and reliability are described in depth. The materials and stimuli used are thoroughly discussed, laying out the conceptual and practical decisions behind their design. Experiment 2 studied language production with the aim to investigate eye movements during conceptual representation, i.e., message generation, when participants intend to speak about motion events. The design and procedure were introduced, with particular focus on the temporal segmentation of the trials into cognitively relevant processing epochs via button presses and speech onsets.

# 4. Analysis

This chapter outlines the analysis procedure, commencing with the details of preprocessing and cleaning applied to the raw data. The calculation of the dependent variables — saccadic travel distances (x/y) and saccade rate — is then explained. Additional data transformation and cleaning procedures are described in the respective sections of each experiment. For each experiment in turn, the hypotheses are operationalized, followed by a description of the specific statistical analysis methods (linear mixed modeling) used to test them. The results of these analyses are presented for each hypothesis.

The first hypothesis of Experiment 1 corresponded to the overarching research question of whether situated simulations in conceptual processing systematically manifest in eye movement patterns, both during non-verbal and verbal event comprehension. Hypothesis 2 examined whether these eye movement patterns were related to the vividness with which participants experienced mental imagery of the events, based on the assumption that stronger simulations – represented by proclivity to experience mental imagery – would be associated with increased oculomotor activity. Hypothesis 3 explored whether eye movement patterns differed between the encoding and recall phases, given that these phases likely involve different weightings of attentional resources to bottom-up versus top-down information flows.

The primary concern of Experiment 2 was to study whether motion direction effects on eye movements also emerge when participants verbally describe such events (Hypothesis 4). Based on the results, further analyses were necessary, as the strongest influence on eye movements originated from the respective processing epoch. Hypothesis 5 examined this finding in more detail by comparing pre-articulation speech planning epochs with the stimulus-replay epoch. In this context, saccade rate was analyzed in conjunction with travel distance to better understand patterns of travel distance variation. An additional hypothesis explored whether eye movement effects in these epochs depended on explicit verbalizations of motion direction by the participants. That is, when directionality in the motion event representation was so salient during message generation that it was verbalized inevitably. To investigate this, Hypothesis 6 compared subsets of trials where these spatial components were explicitly verbalized with those where it was left implicit.

## 4.1. Analysis software

Data processing and analysis was carried out in RStudio (v. 2024.01, *R Development Core Team*, 2008). First, raw gaze data was preprocessed to retrieve the dependent measures *travel distance (x/y)* and *saccade rate*. The dependent variables were analyzed with linear mixed-effects regression modeling (package *lme4* v. 1.1-35.3, Bates et al., 2015) due to the hierarchical nature of the data. Model summaries were generated with package *jtools* (v. 2.2.0, Long, 2022) and visualized with *sjPlot* (v. 2.8.16, Lüdecke, 2024) and *ggplot2* (v. 3.5, Wickham, 2016). Overall significance of regression coefficients was calculated with Wald tests (*MASS* v. 7.3-53, Venables & Ripley, 2002). Beyond visualizations of residual characteristics, inspections of model fit included tests for multicollinearity of predictor variables (*performance, v. 0.11*, Lüdecke et al., 2021). Fixed-effects structure was kept largely identical, drawing on the experimental conditions and hypothesized associations. Depending on the specific hypothesis, certain predictors or subsets of data were excluded.

## 4.2. Data processing

### 4.2.1. Saccade detection

As described in Chapter 3.2, the eye movement type under investigation are saccades. Eye-tracking data was exported using the hardware manufacturer's native software *SR Research DataViewer*. *DataViewer* exports user-friendly gaze event reports that list fixations and saccades, interrupted by blinks and otherwise lost samples in consecutive order at a resolution of 1000 Hz. Unexpectedly, the software's integrated event detection algorithm (velocity- and acceleration-based, like the one by Engbert & Kliegl, 2003; see SR Research Ltd., 2022: 101-104) is rather limited when recorded eye samples are noisy, yielding, for instance, small saccades with physiologically implausible durations (> 200 ms).

To circumvent this issue, raw gaze samples of both eyes were exported, and binocular saccade detection was performed with package *saccadR* (v.0.1.3; Pastukhov, 2022). Raw gaze coordinates (in pixel units) were transformed into degrees of visual angle. *saccadR* extracted all potential saccades from the continuous, raw data stream using three different detection algorithms (each described in Engbert & Kliegl, 2003; Otero-Millan et al., 2014; Nyström & Holmqvist 2010). Whether thus detected

saccade candidates were kept in the algorithm's results output depended on a majority vote-threshold.[62] For the present analyses, this threshold was set to two, that is, the detected saccades had resulted as candidates according to two of the three detection algorithms. In contrast to saccade detection in *DataViewer*, the *saccadR* package thus maximized data validity, increasing the likelihood that the saccade data consisted of genuine saccades and less of technical artefacts introduced by noise or data loss.

The *saccadR* reports contained binocular saccades for Experiment 1 (n=49231, distributed among the non-verbal, n=25027, and verbal condition, n=24204) and Experiment 2 (n=30738). All saccades were then labeled as belonging to distinct trial epochs based on their onset timestamp (cf. trial design in Chapter 3). Saccade-less trials were thus excluded from analysis as saccadR does not output empty trials. The resulting data set is henceforth referred to as saccade data.

### 4.2.2. Saccade data cleaning

Saccade data cleaning and processing consisted of several steps. First, extreme population outliers were removed. This excluded saccades whose raw amplitudes in both x- and y-axes exceeded values beyond three standard deviations (SD) above the respective mean population amplitudes. The +3 SD-threshold roughly compares to amplitudes of 10° horizontal and 7° vertical, measured on the whole sample. Note that on the display computer screen, the grey square, within which participants were instructed to maintain gaze, subtended 13.3° squared visual angle. It is likely that any saccade above 3 SD moved gaze outside of the square's boundaries and is thus unlikely to have resulted from focused attention to the task. Moreover, such large saccades were outliers with respect to the calculation of the dependent variables (see below), and it thus became necessary to remove them from the data of Experiment 1 (1996 of 49231, ~ 4%) and Experiment 2[63].

62 »Each method votes whether a given sample belongs to a saccade. Next, saccades are identified via a majority vote using the *vote_threshold* parameter, as well as a minimum duration and minimal temporal separation criteria.« (Pasthukov, 2022: package 'saccadr' documentation, p. 5)

63 For details on data processing in Experiment 2, see Ch. 4.5 below.

### 4.3. Dependent variables

The critical measures of interest, travel distance and saccade rate, were calculated from the saccade data. Both measures were calculated per trial (Experiment 1) or based on data from relevant epochs exclusively (Experiment 2).

***Travel distance(x/y)*** is the cumulative sum of all absolute saccadic amplitudes per dimensional axis. Every detected saccade is specified with respect to how far it moved the eye (in visual degrees) on the horizontal x-axis and, at the same time, on the vertical y-axis. While saccadic amplitude would refer to the Euclidean distance between the eye's pre- and post-saccadic resting location in estimated gaze coordinates (both x and y positions combined), thus capturing objective amplitude of eye movement that is independent of direction, travel distance aggregates how much distance the eyes covered with saccades in each axis per trial. The gaze data depicted in Figure 4-1 below shows the continuous estimated gaze positions (in blue) of a participant in the critical segment of one trial in Exp. 1. In the relevant, stimulus-specific timespan of 3584 ms, two saccades were measured (left image)[64], yielding a rate of 0.56 saccades per second ((2 saccades ÷ 3584) x 1000 ms). The axis-specific amplitudes of these two saccades (right image) were then summed to yield trial-specific travel distances, i.e., $x_1+x_2$ equals travel distance(x) and $y_1+y_2$ equals travel distance(y).

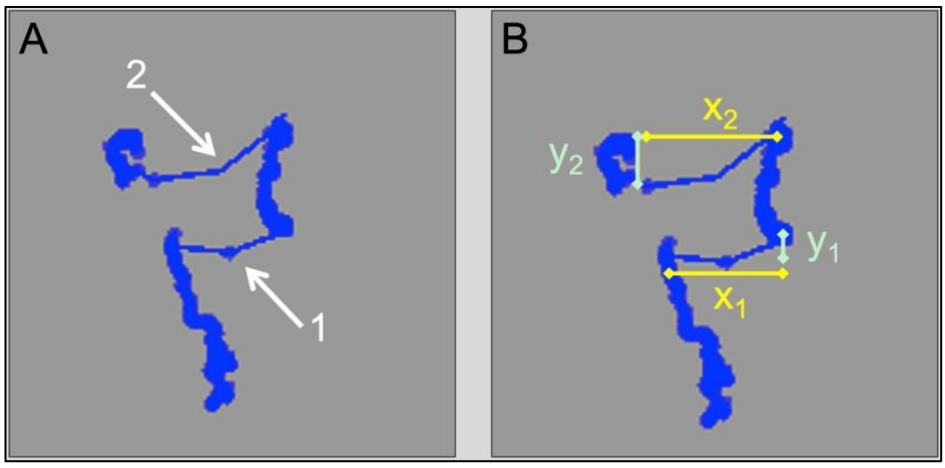


***Figure 4-1****:* Calculation of travel distance(x/y). Both images depict the two-dimensional raw gaze positions of a participant in the critical epoch of a single trial in blue. In panel A, white arrows indicate the two saccadic eye movements. Panel B sketches the logic of travel distance calculation. The spatial displacement of the gaze position through saccades was summed independently per axis, yielding the cumulative magnitude of saccadic gaze shifts in each x and y.

[64] The otherwise vertical displacement of the estimated gaze position is caused by ocular drift, a slow, non-saccadic fixational eye movement.

As Figure 4-1 illustrates, saccade trajectories are seldom plainly horizontal nor vertical, let alone perfectly linear, but this cumulative variable treats them as such — why, then, use travel distance as the dependent variable and not saccade amplitudes (i.e., Euclidean distance between its start and end coordinates) combined with respective direction angles? First, saccade amplitudes vary strongly among participants. Some participants kept their gaze relatively still and exhibited saccades with comparatively small amplitudes. Valid portions of their saccade data would be removed as below threshold population outliers. Second, Euclidean distance collapses both dimensions and would represent travel distance as a single measure of magnitude. Though reduction is attractive for statistical calculations, Euclidean distance is insufficient for the present analysis because it is agnostic about the extension of eye movements in a particular direction. Saccade angles, as an additional measure to Euclidean distances, could solve this, yet again introducing two dependent signals that require independent modeling. In fact, travel distances per directional axis also represent two measures which originated from a single phenomenon, since each saccade always changes pupil-center location in both x and y. However, measuring the full extension of all saccades per trial in the horizontal as well as the vertical axis aligns neatly with the operationalization that the *movement direction* condition of stimuli influences the spatial extension of eye movements, be it horizontal (x) or vertical (y), or both. As a cumulative measure, travel distance can directly capture such trends.

***Saccade rate*** is a single frequency measure computed by dividing the number of saccades per trial by the corresponding trial duration, yielding saccades per second. Analyses of rate were included to help explain variance in travel distances. It is likely that travel distance is positively correlated with rate, in that higher saccade rate brings about larger cumulative travel distances. Interpreting travel distance findings without considering saccade rate could lead to erroneous conclusions, since disregarding how it covaries with rate may inflate the explanatory power ascribed to travel distance as the signal.

When both measures are interpreted together, the term used here is ***oculomotor activity***. Travel distance and saccade rate are both derived from the same underlying saccade occurrences. However, while travel distance is a measure of spatial displacement, saccade rate is based on time. Analyzing both metrics together provides

insight into, for instance, whether observed effects on travel distance result from a high frequency of small saccades or from a few large saccades executed in response to specific stimuli. This approach allows a more precise disentanglement of these interdependent measures and may isolate the effects attributable to the experimental conditions. Oculomotor activity is analyzed in Hypotheses H2, H3, H5, H6.

## 4.4. Experiment 1

Experiment 1 yielded data from a memory task with auditory stimuli that referred to different motion and non-motion events. In two sessions approximately 4 weeks apart, 42 participants were asked to encode and recall non-verbal environmental sounds (Session 1) and short, spoken descriptions of these sounds (Session 2). Each session was split into 4 blocks with 2x12 trials each, resulting in a total of 96 trials. The overarching goal of Experiment 1 was to examine eye movements during event comprehension in the non-verbal versus verbal auditory modality.

### 4.4.1. Trial segmentation

Of each trial, two consecutive temporal segments were isolated for analysis (refer to Chapter 3.3.5.2 for details on trial design). First, the period of audio replay, that is, while the stimulus is being played to the participants. They sense and perceive the stimulus and may already recognize the unfolding event. The second period were 2000 ms of silence and immediately followed the offset of audio replay. This gives participants time for further cognitive processing and memory encoding of the stimulus. The experiment was programmed so that each trial contained these two crucial periods. The calculation of the travel distances and saccade rates (cf. Ch. 4.3) was based on saccade data from these periods. After these periods, routing of trial procedure would vary, playing either a sensory question or the next item, all of which would interrupt focused processing of the current stimulus. On average, the two periods amounted to recorded durations of 4325 ± 594 ms (non-verbal) and 3511 ± 242 ms (verbal condition) (see Table 4-2 below).

### 4.4.2. Discarding invalid trials

For further experiment-specific data cleaning, the dependent variables were stored in separate data sets, depending on the experimental session (non-verbal vs. verbal) and type of response variable (travel distance(x/y) or saccade rate).

In a first step, all recall trials with button presses were discarded (n=725). Complying with task instructions (cf. Ch. 3.3.5.2), participants pressed a button as soon as they recognized a stimulus as novel and thus indicated that they had not heard the sound before. This keypress decision may be correct given that the stimulus had not been played in the encoding phase. It is uncertain whether cognitive processing of such novel stimuli unfolds with the same depth as that of repeated stimuli. After all, the decision to press required a 'no', which was preceded by the realization that a stimulus had simply not been heard. This can be concluded from memory and does not necessarily require in-depth semantic categorization of the current stimulus event. Accordingly, it is uncertain whether any deeper processing occurred before a keypress.

Alternatively, participants' button presses may be incorrect because they falsely recognized a repeated stimulus as novel. In this scenario, participants' button presses may be indicative either of errors in memory performance or of deviant categorizations of stimuli in the encoding phase (i.e., convinced to have heard a different sound before). Either way, it remains uncertain to which degree the cognitive processes that resulted in the keypress were associated with the event semantics of the stimulus situations, rendering the trial invalid. It is crucial for the analysis that participants processed the intended events in a non-superficial manner on all encounters, generating sufficiently comparable situated simulations.[65]

### 4.4.3. Normalizing data for linear mixed-effects modeling

Once trials with button presses had been excluded, further outlier removal was necessary to prepare statistical modeling. Since the distribution of the critical variables was not Gaussian and exhibited a strong right skew, travel distance and rate were logarithmized and cleaned off extreme outliers beyond ±3 SD of the log-mean.

[65] Despite this theoretical logic, there was a methodological motive to discard trials with button presses. In the recall phase, a button press terminated the present trial and fast forwarded to the next one, thereby shortening the duration of analysis segments and inter-stimulus intervals, rendering them incomparable between task phases. This additional variance in trial durations would have to be considered in statistical modeling, possibly weakening statistical power and affecting interpretability.

For saccade rate data, a stricter threshold of ±2 SD was chosen because ±3 SD would have included extreme rates of 1 saccade in 10 seconds (practically not shifting gaze) or 7 saccades per second (suggestive of signal distortion). A threshold of ±2 SD translated to at least 1 saccade in a timespan of 4 seconds and at most 4 saccades per second, which seemed plausible in terms of the instruction to keep gaze on the blank screen display and regarding typical oculomotor patterns (cf. Land, 2019).

Still failing to meet distributional assumptions of linear mixed modeling, travel distances and saccade rate were each transformed with a Box-Cox power transformation (package *MASS* v. 7.3-53, Venables & Ripley, 2002). The Box-Cox transformation is a widely used mathematical tool to produce a close to normally distributed variable with homoscedastic variance (Daimon, 2010: 177). Travel distances and rate were transformed independently of each other and by session.

| | Stimulus modality | Number of trials | Trial duration (± SD) | Minimum–maximum duration | Avg. trials per subject (total = 96) |
|---|---|---|---|---|---|
| Travel distance(x/y) | non-verbal | 3011 | 4325 (± 594) | 2740–5915 | 72 |
| | verbal | 3052 | 3511 (± 242) | 3169–3922 | 73 |
| Saccade rate | non-verbal | 2839 | 4308 (± 600) | 2740–5915 | 68 |
| | verbal | 2897 | 3502 (± 243) | 3169–3991 | 69 |

***Table 4-2**:* Descriptive statistics of trials used for statistical modeling in Experiment 1. The number of trials counts the trials after outlier removal. Durations are given in milliseconds. Average trials per subject illustrates that about 24 of the 96 trials were removed as containing outlier or invalid data. For distribution of trials across conditions, refer to Appendix B1.

### 4.4.4. Standardizing trial durations

Trial durations were significantly different between stimulus modalities (see Ch. 3.3.4.3). Environmental sounds have different play durations in all stimulus conditions (cf. top-left panel in Fig. 4-3 below). Stimuli referring to horizontal motion were longest, outlasting vertical motion stimuli and non-motion stimuli to significant degrees. Coincidentally, this imbalance also affected the verbal stimuli, though to a lesser extent. Given that there is more time available to execute saccades, it is likely that the longer the trial, the more travel distance accumulates. This may impact statistical analysis in that stimulus conditions might significantly affect travel distances due to the different trial durations associated with the condition and not because of the movement direction condition itself. To control for this potential relationship, trial duration was included as

a fixed effect in the statistical models. To avoid multicollinearity between fixed-effects terms *trial duration* and *movement direction condition*, trial durations were z-scored by condition, so that their distributions were centered to means of zero and ± 1 SD (see Figure 4-3, Row B).

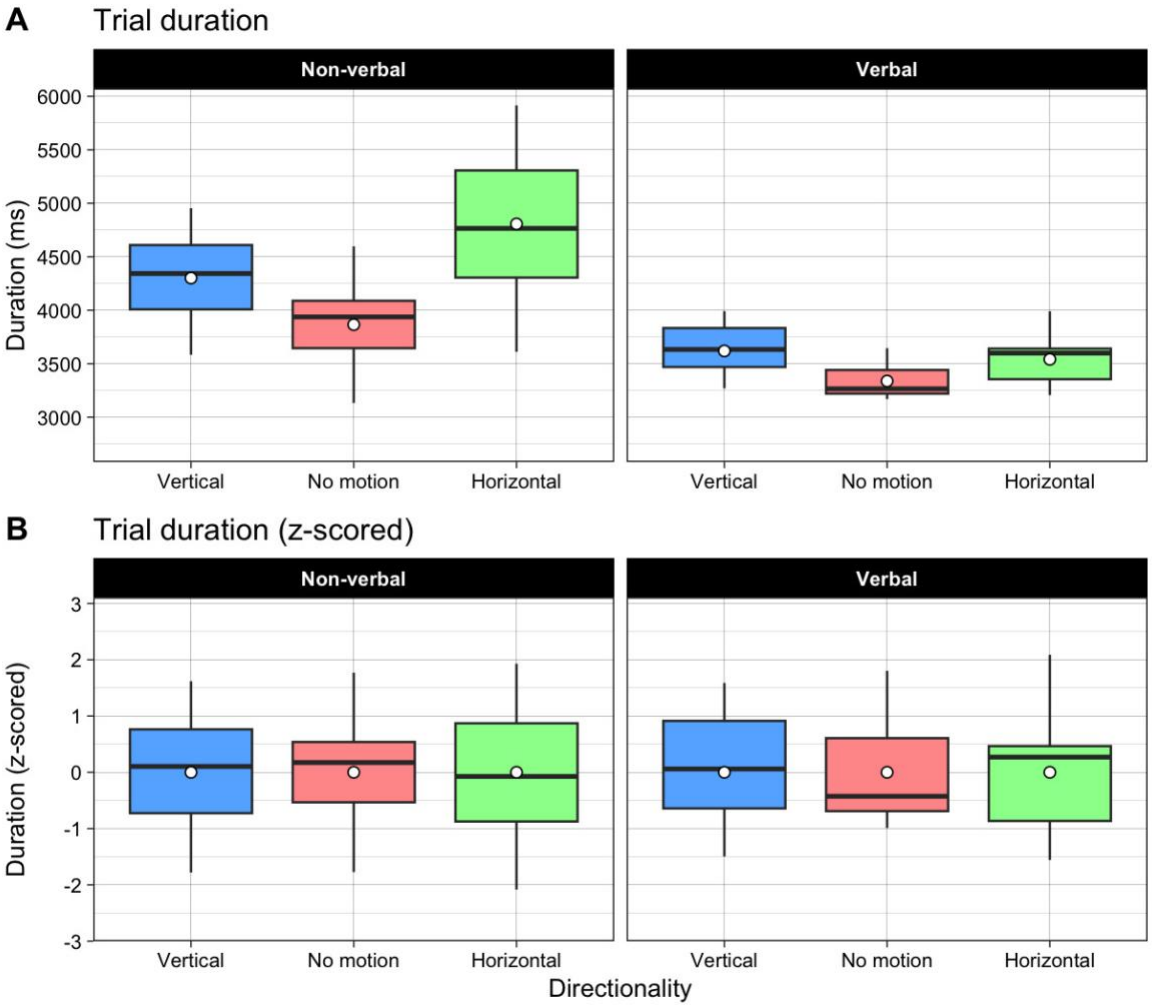


***Figure 4-3**:* Trial durations by movement direction and stimulus modality. Panel A displays raw trial durations between conditions (x-axis) in milliseconds, while Panel B presents z-scored durations. White dots indicate the mean. Panel B illustrates that z-scored trial durations have comparable central tendencies across movement direction conditions and stimulus modalities.

### 4.4.5. Analysis procedure

Mixed-effects linear regression targeted the dependent variables *travel distance(x)*, *travel distance(y)* and *saccade rate* as continuous measures. Predictors included the *movement direction* condition (3 levels: vertical, horizontal, no-motion), participants' self-reported ratings of *visualization intensity* (ordinal factor: 2–5), *task phase* (2 levels: encoding vs. recall) and *stimulus modality* (2 levels: non-verbal vs. verbal). Trial

duration (z-scored continuous numeric) was included as a further predictor but intended as a control variable. Considering that a substantial amount of variance could be explained by participants' individual eye movement behavior, a grouping variable *subject* with random intercept was selected. Additionally, given that eye movement rate varies between participants but influences the accumulated travel distance, a subject-and-trial-specific *saccade rate* was integrated as a random slope for *subject*. The random slope acknowledges that each subject exhibits a unique eye movement rate per trial and that the effect of this rate on travel distance varies across subjects. It thereby reduces the risk of attributing all explanatory power to experimental conditions without accounting for associated covariates. Mixed-effects linear regression is a powerful tool to analyze data with such hierarchical characteristics in that it models the diverse factors in one calculation while accounting for the dependence and variance associated the grouping levels (e.g., participants and items) (Meteyard & Davies, 2020).

For easier interpretation of coefficients, the categorical predictors were dummy-coded. Dummy-coding means that one level of the independent variable (e.g., the control condition) is used as a reference level against which the other levels are contrasted (Brehm & Alday, 2022). The reference levels were set according to the hypothesized contrasts. Effects of critical stimulus conditions (horizontal and vertical motion) were calculated with respect to each other and the control condition (non-motion events). Regarding the self-reported visualization intensity, a rating of 4 was selected as the reference level since it made up most data. Out of the two task phases, the recall phase was contrasted with the encoding phase since the latter is more homogenous and contains more data. In the comparison of stimulus modality, the data from the verbal stimuli was compared to that of the non-verbal data.

Model syntax was determined through stepwise reduction of a maximal model (Barr et al., 2013) with comparisons of each model's AIC performance and residual characteristics. Having met stochastic requirements (convergence criterion), ultimate selection of model syntax was guided by its alignment with the hypotheses, opting for the most plausible configuration. Model fits were assessed by visualizing the distribution of the residuals and their relationship to the fitted values.

Depending on the degree of detail in post-hoc-examination of the hypotheses, certain predictors were excluded, but full model syntax followed this formula:

| | |
|---|---|
| *travel distance(x/y)* | *~ movement direction + visualization + task phase + trial duration + stimulus modality* + (1+*rate*\|*subject*) |
| *saccade rate* | *~ movement direction + visualization + task phase + trial duration + stimulus modality + trial number* + (1\|*subject*) |

### 4.4.6. Formal hypotheses

The main objective of this experiment is to examine whether oculomotor activity is affected by semantic properties of space in motion events when subjects are tasked with recognizing, encoding, and recalling auditory event stimuli as environmental sounds (non-verbal) and spoken descriptions (verbal). Three hypotheses were formulated:

H1) Critical stimuli increase travel distance in the axis parallel to movement direction. Motion events with horizontal trajectories increase travel distance(x) and motion events with vertical trajectories increase travel distance(y) as opposed to non-motion events and those with orthogonal trajectories.

$H_1$: $\text{Travel distance(x)}_{\text{HORIZONTAL}} > \text{Travel distance(x)}_{\text{VERTICAL AND NO-MOTION}}$

$\text{Travel distance(y)}_{\text{VERTICAL}} > \text{Travel distance(y)}_{\text{HORIZONTAL AND NO-MOTION}}$

$H_0$: $\text{Travel distance(x)}_{\text{HORIZONTAL}} \leq \text{Travel distance(x)}_{\text{VERTICAL AND NO-MOTION}}$

$\text{Travel distance(y)}_{\text{VERTICAL}} \leq \text{Travel distance(y)}_{\text{HORIZONTAL AND NO-MOTION}}$

H2) Oculomotor activity exhibits a positive relationship with self-reported visualization intensity.

$H_1$: $\text{Travel distance}_{(X/Y)}$ & rate ~ visualization > 0

$H_0$: $\text{Travel distance}_{(X/Y)}$ & rate ~ visualization ≤ 0

H3) Exploratory: Oculomotor activity differs between encoding and recall phases.

$H_1$: $\text{Travel distance}_{(X/Y)}$ & rate ~ task ≠ 0

$H_0$: $\text{Travel distance}_{(X/Y)}$ & rate ~ task = 0

### 4.4.7. Results

Results listed in the following two tables will be described in detail in the sections for Hypotheses 1, 2 and 3. Table 4-4 lists main effects from the full model summaries. For comprehensive summary tables, refer to Appendix B2.

| | **Travel distance(x)** | | | **Travel distance(y)** | | | **Saccade rate** | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Estimate* | *CI* | *p* | *Estimate* | *CI* | *p* | *Estimate* | *CI* | *p* |
| (Intercept) | 1.40 | 1.17 – 1.62 | <0.001 | 0.28 | 0.08 – 0.48 | 0.007 | -0.09 | -0.22 – 0.03 | 0.14 |
| Movement direction [vertical] | 0.16 | 0.12 – 0.20 | <0.001 | 0.12 | 0.08 – 0.17 | <0.001 | 0.02 | -0.01 – 0.05 | 0.23 |
| Movement direction [horizontal] | 0.28 | 0.24 – 0.33 | <0.001 | 0.19 | 0.14 – 0.24 | <0.001 | 0.01 | -0.02 – 0.04 | 0.47 |
| Visualization intensity [5] | -0.37 | -0.75 – 0.00 | 0.05 | -0.10 | -0.44 – 0.24 | 0.559 | -0.20 | -0.41 – 0.01 | 0.06 |
| Visualization intensity [3] | 0.27 | -0.16 – 0.71 | 0.22 | 0.16 | -0.24 – 0.55 | 0.435 | -0.12 | -0.36 – 0.12 | 0.32 |
| Visualization intensity [2] | 0.49 | -0.06 – 1.04 | 0.08 | 0.34 | -0.15 – 0.84 | 0.172 | 0.12 | -0.18 – 0.42 | 0.44 |
| Trial duration (z-scored) | 0.13 | 0.11 – 0.15 | <0.001 | 0.10 | 0.08 – 0.12 | <0.001 | 0.01 | -0.00 – 0.02 | 0.06 |
| Task phase [recall] | -0.21 | -0.25 – -0.17 | <0.001 | -0.09 | -0.14 – -0.05 | <0.001 | 0.32 | 0.30 – 0.35 | <0.001 |
| Stimulus modality [verbal] | -0.71 | -0.74 – -0.67 | <0.001 | -0.50 | -0.54 – -0.46 | <0.001 | 0.13 | 0.11 – 0.16 | <0.001 |
| Trial number | – | – | – | – | – | – | 0.002 | 0.001–0.002 | <0.001 |

***Table 4-4**:* Linear mixed model results for Experiment 1. Estimates, 95% confidence intervals (*CI*), and p-values are reported for three dependent variables: travel distance in the x-axis, travel distance in the y-axis, and saccade rate. Predictors include movement direction, visualization intensity, trial duration (z-scored), task phase, and stimulus modality. The reference levels for categorical predictors are movement direction [no motion], visualization intensity [4], task phase [encoding], and stimulus modality [non-verbal]. Random effects and model fit are listed in Appendix B2.

| | **Travel distance(x)** | | **Travel distance(y)** | | **Saccade rate** | |
|---|---|---|---|---|---|---|
| | Estimate (*df*) | *p* | Estimate (*df*) | *p* | Estimate (*df*) | *p* |
| Movement direction | $\chi^2(2) = 161$ | < 0.001 | $\chi^2(2) = 62$ | < 0.001 | $\chi^2(2) = 1.47$ | 0.48 |
| Visualization intensity | $\chi^2(3) = 11.6$ | 0.009 | $\chi^2(3) = 3.4$ | 0.34 | $\chi^2(3) = 5.7$ | 0.13 |
| Task phase | $\chi^2(1) = 110$ | < 0.001 | $\chi^2(1) = 18.4$ | < 0.001 | $\chi^2(1) = 619$ | < 0.001 |
| Stimulus modality | $\chi^2(1) = 1530$ | < 0.001 | $\chi^2(1) = 641$ | < 0.001 | $\chi^2(1) = 116$ | < 0.001 |
| Trial dur. (z-scored) | M = 0.13 | < 0.001 | M = 0.10 | < 0.001 | M = 0.01 | 0.06 |

***Table 4-5**:* Wald-test results show significant associations of predictors with response variables in Experiment 1. For categorical predictors, $\chi^2$ (chi square) represents the significance of categorical differences. For continuous predictors, M represents the mean estimated effect size on the response variable.

#### 4.4.7.1. Hypothesis 1

Statistical modeling confirmed main effects of movement direction on travel distance, both in the non-verbal and verbal condition and across task phases, controlled for trial duration and participants' visualization ratings. Effects on the two dependent variables travel distance in the x- and y-axes were modelled separately.

Processing horizontal stimuli led to an increase of travel distance(x) (Est. = 0.28, $p < 0.001$) as opposed to non-motion stimuli. Similarly, vertical stimuli were also associated with an increase in horizontal travel distance (Est. = 0.16, $p < 0.001$), though to a lesser degree. Beyond these dimensional levels, movement direction was a significant predictor for changes in travel distance(x) overall ($\chi^2(2) = 161$, $p < 0.001$).

Participants exhibited larger vertical travel distance(y) when exposed to vertical stimuli (Est. = 0.12, $p < 0.001$) than when they heard non-motion stimuli. Horizontal stimuli, however, led to even larger increases in vertical travel distance (Est. = 0.19, $p < 0.001$). Assessing motion direction with respect to the other predictors, a Wald-test confirmed that its collective effects on travel distance(y) are significantly different from zero ($\chi^2(2) = 62$, $p < 0.001$).

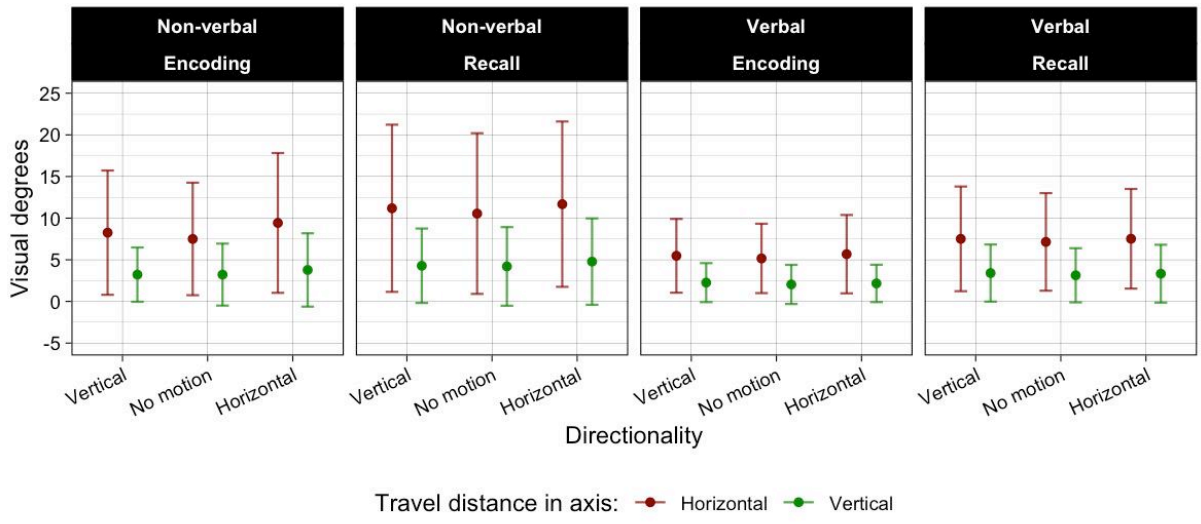


***Figure 4-6**:* Travel distances (horizontal = red) and (vertical = green) in visual degrees (y-axis) and by movement direction conditions (x-axis) in both modalities and task phases. Error bars show ±1 SD. As opposed to stimuli without motion, vertical and horizontal stimuli show increases of travel distance, especially in the horizontal axis.

In view of these results, Hypothesis 1 can be confirmed in the general sense that both horizontal and vertical motion events enlarge travel distances in the respective motion axes. The claim that this increase would affect travel distance more strongly in the parallel eye movement axis as opposed to the orthogonal axis is only partially

supported by statistical evidence. Despite the increase of horizontal travel distance being largest for horizontal motion events, the corresponding trend for vertical motion events was not present in the data. In fact, the largest increase of vertical travel distance was measured in events where movement direction was horizontal. Consequently, the specific hypothesis that movement direction triggers larger eye movements in the corresponding directional axis can only be confirmed for the horizontal plane.

Given that these findings for travel distance were already controlled for saccade rate with the grouping term, effects of motion direction on saccade rate were not of primary concern for Hypothesis 1. For sake of completeness, however, results show no significant influence ($\chi^2(2) = 1.47$, $p = 0.48$), neither for the horizontal (Est. = 0.01, $p = 0.47$) nor vertical stimulus condition (Est. = 0.02, $p = 0.23$). No effect of movement direction on saccade rate was detected by the model.

Tables 4-4 and 4-5 summarized further predictors affecting both horizontal and vertical travel distance, confirming that travel distances are partly affected by participants' degree of visualization (cf. Hypothesis 2) and vary between stimulus modalities, task phases (cf. Hypothesis 3), and with trial duration.

In summary, participants exhibited larger travel distances in both axes when they processed motion events as opposed to non-motion events, regardless of whether stimuli were presented as environmental sounds or verbal descriptions, independent of participants' self-reported visualization intensity, and irrespective of whether they were tasked with encoding or recalling them. This suggests that processing a motion event with discrete direction results in a larger spatial dispersion of non-visual gaze.

#### 4.4.7.2. Hypothesis 2

The second hypothesis examined the relationship between participants' self-reported intensity of experience of mental images and their oculomotor activity. After the first session, participants rated the vividness of their mental visualizations on a Likert scale of 1 to 5 (see Table 4-7 below).

| | | Self-reported visualization intensity | | | | |
|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **5** |
| Exp. 1 | Subjects (n=42) | – | 4 | 7 | 20 | 11 |
| | Proportion of sample | – | 9.5% | 16.7% | 47.6% | 26.2% |
| Exp. 2 | Subjects (n=42) | – | 3 | 9 | 14 | 16 |
| | Proportion of sample | – | 7.1% | 21.4% | 33.3% | 38.1% |

***Table 4-7**:* Distribution of self-ratings of visualization intensity in Experiment 1 and 2. None of the subjects gave a rating of 1. Ratings were given at the end of each experimental session.

The crucial measure per trial, oculomotor activity, is expressed by the saccadic travel distances in the x- and y-axes in relation to the number of saccades it took to amount to such travel distances. These three response variables were examined with separate models, which are summarized in Table 4-4 above and reported below.

Visualization intensity was significantly associated with changes in horizontal travel distance ($\chi^2(3)$[66] = 11.6, $p < 0.01$). Compared to the majority, who gave a rating of 4, subjects with a rating of 2 moved their eyes further in the x-axis (Est. = 0.49, $p$ = 0.08) than participants who gave a rating of 5 (Est. = -0.37, $p$ = 0.05). The marginally significant p-values reported here may be due to the differences in sample sizes between the groups (see Table 4-7 above), though revealing a trend that visualization groups differ from each other with respect to their horizontal travel distances. The estimates imply that this relationship is inverse. Participants who reported experiencing strong mental imagery moved their eyes across smaller horizontal distances whereas those who reported weak visualizations seem to have covered larger distances with their eye movements.

The identical model for vertical travel distance revealed no significant effects of visualization ($\chi^2(3) = 3.38$, $p = 0.34$). Detailed examinations of session and task phase

[66] All participants reported visualizations between 2 and 5, which results in three degrees of freedom.

subsets did not yield correlations between changes in travel distance(y) and levels of visualization either. Neither of the models showed significant correlations of individual visualization levels with changes in travel distance(y). No evidence was found that eye movement in the vertical axis was affected by visualization intensity.

Saccade rate is inextricably linked to travel distance in that the occurrence of a saccade generates travel distance. Hence, given that there was a significant overall association of visualization with travel distance(x), a similar relationship with rate is expected. Contrasting with theoretical predictions, saccade rate was not significantly altered by self-reported visualization intensity ($\chi^2(3) = 5.7$, $p = 0.13$).

A closer examination revealed (cf. Appendix B3 for results table), however, that rate was not affected in the recall phases ($\chi^2(3) = 2.8$, $p = 0.42$) but in the encoding phases ($\chi^2(3) = 10$, $p < 0.01$). In fact, strongly visualizing participants exhibited a reduction of saccade rate (Est. = -0.22, $p = 0.01$) whereas participants with visualization ratings of 2 showed a slight but insignificant increase of rate (Est. = 0.11, $p = 0.38$). This may be related to the finding for horizontal travel distance insofar that since participants who experienced mental imagery to different degrees exhibited covariant cumulative travel distances, so saccade rate may be the oculomotor signal associated with subjects' relying on or generating of mental images for comprehension (see Chapter 5.2.4).

Hypothesis 2 stated that self-reported visualization intensity was positively correlated with oculomotor activity, increasing both travel distances and saccade rate. The present findings partly suggest a correlation, but for the inverse direction: Horizontal travel distance varied with visualization ratings, enlarging when imagery was experienced as weak and vice versa. No significant associations were found for vertical travel distance and rate overall. However, a filtering of the modeled data for encoding trials only revealed that rate was affected in a similar pattern. Fewer saccades were executed in encoding phases by participants who reported strong imagery — they kept their eyes rather still.

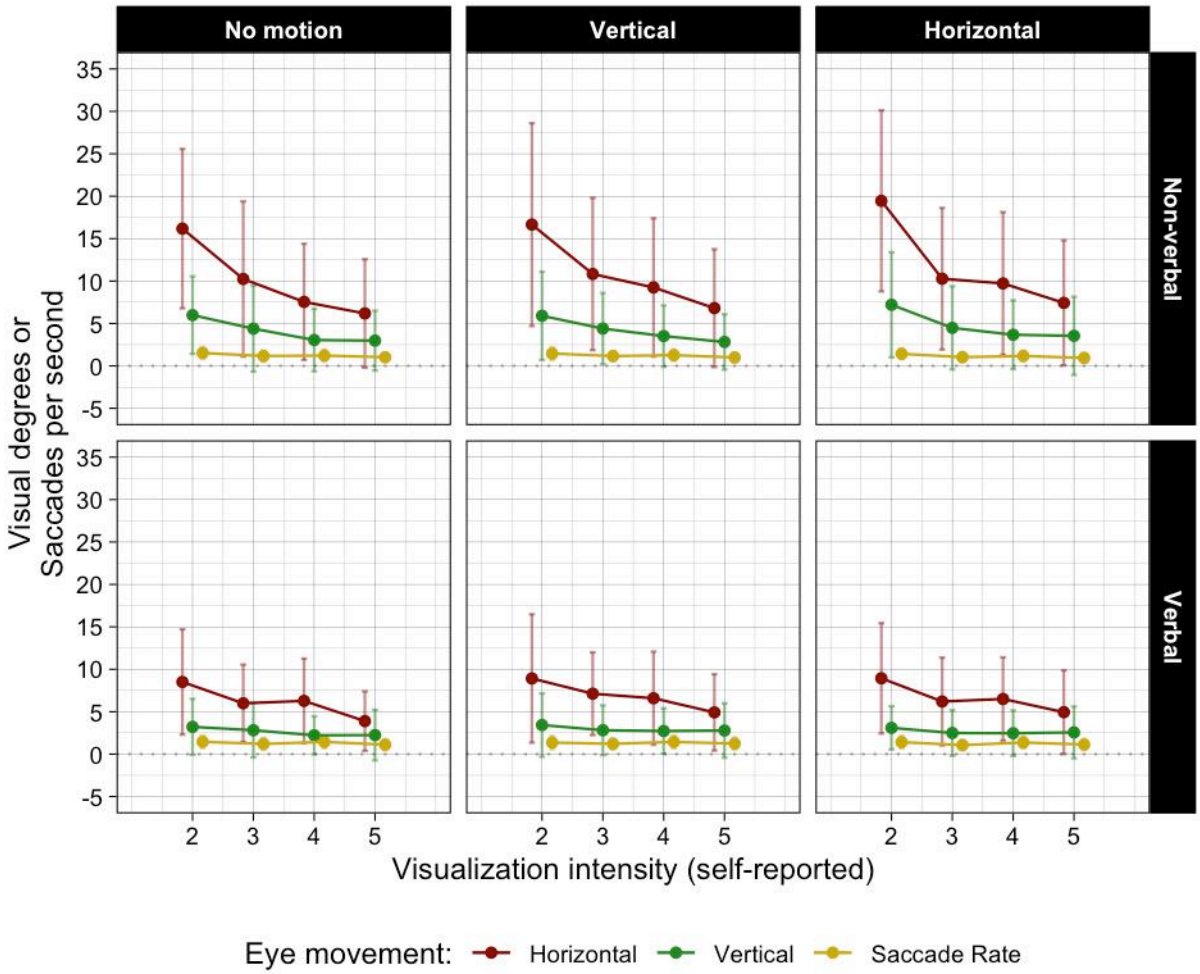


***Figure 4-8**:* Travel distances (horizontal = red) and (vertical = green) in visual degrees, as well as saccade rate (yellow) in saccades per second, aggregated per group of visualizers (x-axis), by stimulus movement direction (columns) and modality (rows). Error bars show ±1 SD. Travel distances decrease, in both x and y, with increasing visualization intensity.

#### 4.4.7.3. Exploratory Hypothesis 3

The reported differences in eye movements between task phases (see results H1, H2) warrant further exploration. Therefore, the goal of Hypothesis 3 was to investigate whether the distinct task phases impacted participants' oculomotor activity in systematic ways, given that the encoding phase required recognition and memory encoding, whereas, in the recall phase, task compliance mobilized resources mainly for recognition based on working memory.

All three dependent variables were significantly influenced by task phase in both modalities (see Table 4-4 above). In comparison to the encoding phase, travel distance decreased in the recall phase, both in the horizontal (Est. = -0.21, $p < 0.001$) and vertical axes (Est. = -0.09, $p < 0.001$). Saccade rate, on the other hand, increased in

the recall phases (Est. = 0.32, $p < 0.001$), with task phase as the strongest predictor overall ($\chi^2(1) = 619$, $p < 0.001$) (cf. Table 4-5). In other words, although participants initiated more saccades during recall, they seem to have covered comparatively shorter cumulative travel distances with this higher number of saccades.

To examine this statistically, a further mixed model was fit to assess the interaction between saccade rate and task phase in their effects on horizontal travel distance (x) post-hoc (cf. Appendix B4). The model confirmed the assumed interaction. While both rate (Est. = 1.55, $p < 0.001$) and task phase (Est. = 0.25, $p < 0.001$) on their own were significantly associated with increases in travel distance (x), their interaction term (Est. = -0.35, $p < 0.001$) indicated a decrease in travel distance at higher saccade rates during the recall phase. This suggests that the effect of saccade rate on travel distance (x) was weaker during recall than during encoding. In other words, although participants made more saccades during the recall phases, the horizontal travel distance they accumulated per saccade was smaller compared to the travel distance accumulated during the encoding phases.

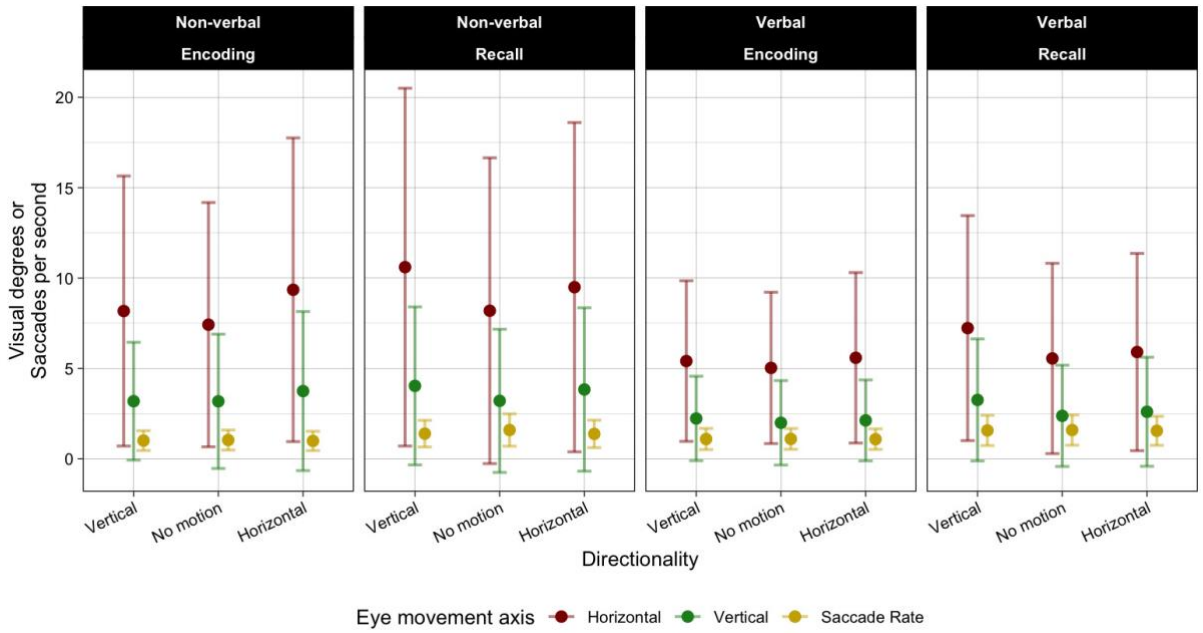


***Figure 4-9**:* Travel distances (horizontal = red) and (vertical = green) in visual degrees, as well as saccade rate (yellow) in saccades per second, by movement direction conditions (x-axis) in both modalities and task phases. Error bars show 1 standard deviation. Saccade rates (yellow) are significantly increased in the recall phases as compared to encoding. While larger saccade rates during encoding are linked to increased travel distances, this association is weaker in the recall phase, where comparatively increased rates do not correspond to equally proportionate travel distance gains and may even show a negative trend. Travel distances do not increase as much due to higher rates in recall as would be expected from the saccade rate x travel distance-relationship in encoding.

Additional findings from the model on saccade rate (cf. Table 4-4) entail significant, but minimal increases of saccade rate due to time spent on experiment (main effect of trial

number (Est. = 0.002, $p < 0.001$) (see Ch. 5.3.4.1), and a slight increase of saccade rate in the verbal condition (Est. = 0.13, $p < 0.001$) as opposed to the non-verbal condition.

### 4.4.8. Summary of the results (Exp. 1)

Experiment 1 investigated whether the comprehension of motion events is accompanied by oculomotor activity that suggests situated simulation as a means of conceptual representation.

While participants memorized and recalled both non-verbal and verbal event stimuli, their eye movements covered larger distances in the motion event condition compared to the non-motion event condition. This larger spatial extension of eye movements was not associated with axial directionality of the motion, contrary to the assumption of Hypothesis 1.

Hypothesis 2 tested whether these eye movements, in particular their overall magnitude and frequency, would be modulated by participants' self-reported intensity of experience of mental images. This relationship was found to be inverse, in that the more vivid the visualizations, the shorter the travel distance (x) and the lower the saccade rate. Effects on saccade rate only existed during encoding phases, not during recall, suggesting that task demands influence oculomotor activity as well.

This motivated a closer look at potential task effects. Hypothesis 3 compared the encoding and recall data and confirmed a strong influence of task phase. In general, travel distance decreased while saccade rate increased. Participants initiated more saccades but covered shorter travel distances during recall.

Altogether, hypothesis testing in Experiment 1 confirmed general effects of movement direction, individual disposition for visualization, and of task requirements on spontaneous, non-visual eye movements.

## 4.5. Experiment 2

The second experiment was integrated in Session 1 and took place shortly after the first experiment. The same 42 participants were asked to listen to the 64 environmental sound stimuli of Experiment 1 again and to verbally describe what they thought they heard. This task was split into 4 blocks with 16 trials each. At the end, participants were again asked to rate the vividness of their internal visualizations during the task.

### 4.5.1. Trial segmentation

Complementing Experiment 1, Experiment 2 focused on the language production process underlying their descriptions of motion vs. non-motion events and the eye movements that accompany this process. Participants were instructed to describe what is happening and press a button when they were ready to do so. This procedure allowed for the segmentation of each trial into distinct, consecutive segments that roughly corresponded to the language production stages of message generation, formulation, and articulation (Levelt, 1989; refer to Chapter 3.3.6.2 for details).

First, in the period of **audio replay**, the stimulus is being played to the participant. The second epoch, termed **pre-button**, refers to a segment of silence between audio offset and participants' behavioral signal that they are ready to speak (keypress). The next epoch, between the keypress-signal and participants' onset of articulation, was termed **pre-voice**. Finally, the **articulation** epoch began as soon as the participants started speaking. In technical terms, this epoch began in the moment when the experimental software detected the voice for the first time. While this last epoch had a fixed duration of 6000 milliseconds, the other epochs have varying durations and numbers of observations (see Table 4-10). This variation was to be expected, since participants were able to control the trial procedure themselves.

These epochs are distinguished to link eye movement patterns to the different cognitive processes shortly before and during language production. In the audio replay epoch, participants likely perceive and recognize the stimulus and begin generating or formulating the message. The pre-button epoch commenced when the stimulus finished playing and ended when participants hit the response button. Since participants were instructed to press the button as soon as possible and were usually fast at doing so during audio replay, data for this second epoch was often missing.

However, when saccades were measured there, they were taken to be related to ongoing processes supporting message generation, since the message is assumed to be the minimal requirement for one to indicate readiness to speak. The third epoch commenced with the keypress and ended with voice onset. Given that readiness to speak would only be possible as early as the first component of the message was given to the formulator, the eye movements registered in this epoch are interpreted as driven by lexical and syntactic encoding. As soon as the last epoch (articulation) began, most processing resources were channeled towards finishing the utterance and monitoring of phonemic realization, at least when responses were short and simple.

### 4.5.2. Discarding invalid trials

Not all trials contain saccade data for all epochs and epoch duration varies within and between subjects (see Table 4-10 below). This is because participants differed in how they solved the verbalization task. Some seemed to have applied the strategy not to press the button before they were ready to speak, thereby extending the pre-button epoch but cutting short the pre-voice epoch (i.e., button-voice-latency). Other participants pressed the button during audio replay and began speaking shortly thereafter, suggesting that message generation and formulation had proceeded to an articulatory plan before the audio replay had finished. Data unavailability aside, this facilitated removal of outliers and noise. Epochs without measurable saccades were simply not output.

Further data exclusion was based on task compliance and targeted trials in which participants did not recognize the stimulus (by saying “no idea”, on 3±2 out of 64 stimuli) or when they did not recognize it on first encounter and requested the audio be replayed. Out of 42 participants, only 19 requested repetitions at all. Two appeared to use repetition strategically given that they repeated 22% of trials, whereas the remaining 17 subjects did so 2–4 times, on average. In total, 135 trials, 5% of all 2688 unique trials, were removed for non-recognizability. From this subset, all remaining trials with repetition requests were removed, even those in which the participant recognized the stimulus eventually (n=45). In 20 trials, no saccades were detected by the *saccadR* algorithm. Lastly, the data from 13 trials was excluded because the voice-

recording duration was cut short due to software malfunction. In total, 2475 valid trials from 42 participants resulted.

### 4.5.3. Normalizing data for linear mixed modeling

Raw data cleaning based on statistical criteria followed the same procedures described in Chapter 4.2 above. Unusually large saccades with amplitudes beyond a population mean amplitude threshold of +3 SD were removed (about 5% of all saccades; with horizontal amplitudes larger than 10.7° and vertical amplitudes beyond 14.8°). Data from two further trials was removed through this procedure. This dataset (n=2473 trials comprising 24353 saccades) was used to compute travel distances and rates per epoch, which were again stored in two separate data sets, one for travel distances and one for rate.

To prepare the data for linear regression, the dependent variables were subject to further standardizing operations. Logarithmized travel distance (x/y) data was stripped off values exceeding ± 3 SD from the respective population log-mean, allowing for a fine-grained exclusion of epochs with unusually small travel distances (41 observations are removed, less than 1% of data). Saccade rates were logarithmized and cleaned off values further than ±2.5 SDs from the population log-mean (corresponding to a rate of at least 1 and at most 5 saccades per second). This different sigma-threshold for rate was determined upon visual inspection of the skewed distribution of the data. A liberal threshold of ±3 SDs would have retained many outliers, whereas a stricter threshold of ±2 would have cut densely distributed observations. Forcing the dependent variables into a normal distribution, Box-Cox power transformations were applied to each of the travel distance measures and saccade rate. Table 4-10 below gives an overview of the data analyzed for Experiment 2.

| | | Audio replay | Pre-button | Pre-voice | Articulation | |
|---|---|---|---|---|---|---|
| | | *Stimulus is playing* | *Stimulus finished playing but no keypress* | *Key pressed but voice not yet detected* | *Voice onset detected* | all epochs |
| **Travel distance(x/y)** | Number of trials | 1922 | 909 | 1262 | 2432 | 6525 |
| | Trial duration (± SD) | 2083 (± 658) | 1241 (± 760) | 913 (± 389) | 6000 | 8445 (± 1781) |
| | Minimum–maximum duration | 621–3914 | 46–3957 | 115–3598 | – | |
| **Saccade rate** | Number of trials | 1929 | 898 | 1279 | 2331 | 6437 |
| | Trial duration (± SD) | 2081 (± 658) | 1262 (± 753) | 914 (± 387) | 6000 | 8337 (± 2016) |
| | Minimum–maximum duration | 621–3914 | 192–3957 | 206–3598 | – | |

***Table 4-10**:* Description of the data used for statistical modeling in Experiment 2. Organized by dependent variables and trial epochs. The number of trials concerns the trials with valid saccades that remained after the outlier removal procedures described in the sections above. Durations are given in milliseconds.

### 4.5.4. Defining *motion event interpretation* as an independent variable

Participants' verbal descriptions were recorded and synchronized with the eye-tracking data. Each trial started a separate voice recording shortly before the stimulus was played. On average, response utterances were 4 ± 2 words long and were categorized as referring to either vertical or horizontal motion events or non-motion events. The data to be analyzed (n=2472) contains 739 responses expressing horizontal motion events, 603 expressing vertical motion, and 1130 non-motion events. This classification was based on various criteria.

First, verbal responses were checked regarding their agreement with the intended stimulus event. By and large, participants recognized the events accurately and described the source event of the environmental sound. If the response did not describe the actual event but still denoted a motion event, the interpretation was labeled as an event of the corresponding directional category. The same holds for cases where the elicited response was imprecise, for instance, with respect to object information that was irrelevant to movement direction, such as *pouring rice in a container* being interpreted as *lentils falling on tile floor*. Both verbalizations align in motion dynamics (*falling)* and idealized directional axis (*down*). If the response was not clearly classifiable as a motion event description, it was labeled as under-specified and processed again in a later step (see below).

Syntactically, participants expressed motion event descriptions largely with noun phrases for the figure and verb phrases whose head referred to movement. Ideal descriptions contained additional satellite adjuncts that profiled spatial components like GROUND, ENDPOINT or DIRECTION. These were mostly directional particles as separable prefixes of the lexical verbs (e.g., *runter-fallen*, to fall down) or prepositional phrases (e.g., *auf den Boden*, to the ground) specifying location or goal, and sometimes both (e.g., *fällt auf den Boden runter*, falls down to the ground). This way, the response was unmistakably interpretable as referring to an entity in translational motion. For example, the normalized description of vertical motion item 1, *jemand geht eine Treppe herauf* (someone is going upstairs), refers to a human entity with NP *jemand* (someone), who moves in the manner of *gehen* (to walk) and in the perspectivized direction of *herauf* (upward[67]), climbing *eine Treppe* (stairs). Another subject described this same item as *es wird in einer Halle ein Ballspiel gespielt* (a ball game is being played in a gym), which implies movement but without concrete directionality or profiling of a moving entity, such that it was classified as a non-motion event. For the horizontal item *jemand schwimmt im Wasser* (someone is swimming in the water), responses like *jemand planscht im Wasser* (someone is splashing in the water) were classified as non-motion events because translational motion of the figure was not implied and axial movement direction was difficult to infer, making it unlikely to have been salient during event construal.

In total, overt reference to both GROUND and PATH/DIRECTION in satellites was registered 190 times, exclusive reference to DIRECTION was counted 210 times and exclusive reference to GROUND in 4 trials. This subset of responses makes up the data examined in Exploratory Hypothesis 6 below (n=1342).

The remaining motion event descriptions (393 verticals + 545 horizontals = 938) lacked explicit verbalization of spatial components, despite clearly referring to events that contain translational motion (e.g., a faucet is running, a helicopter flight, pouring a glass of water, fireworks, someone is playing ping pong); in many cases, motion components were encoded in the predicate verb but not supplemented with constituents denoting GROUND, DIRECTION or ENDPOINT/GOAL. These under-specified responses may be a result of eliciting natural speech under time pressure. How these responses can still justifiably qualify as motion event descriptions is discussed next.

[67] Deictic directional that indicates upward movement toward the speaker.

### *Categorizing under-specified responses*

Under-specified responses were difficult to classify mainly because of two issues. First, verbal responses did not concretely reveal the intended motion event with its figure-related directional dynamics. For example, when the event was recognized but expressed with a different perspectivization: *ein Glas zerbricht* (a glass is breaking) emphasizes the audible resultative state whereas *ein Glas fällt auf den Boden* (a glass is falling to the floor) expresses motion and endpoint but leaves the figure's resultative state implicit. Second, some stimuli were described as integrated into a larger event frame, defocusing the crucial, inherent translatory motion component (e.g., one-word responses, such as *Feuerwerk*).

In both cases, classification of the verbal response hinged upon the salience of the motion component in the environmental sound. When recognition of the inherent motion event was necessary to establish the interpretation expressed by the under-specified description — that is, when the description clearly implied that the subject had interpreted the objectively unfolding sound event correctly, but simply did not express the motion component, the response qualified as being based on a translational motion event — after all, it must have been perceived as one to achieve that interpretation. This decision was made per item and rests on the assumption that not all that is conceptualized breaches the linguistic surface. This assumption aligns with tenets from grounded cognition (Barsalou, 2008), holding that mental simulations may remain unconscious and that we do not become aware of the breadth of our mental representations when engaged in different tasks. These decision processes are illustrated in the remainder of this section.

In contrast to the ideal descriptions (with DIRECTION or GOAL adjuncts), under-specified responses backgrounded the spatial dynamics of the audible entity-in-motion. Take the example of item 29, which plays an excerpt of a ping pong game. What is audible is the bat-hitting and bouncing of the ball on the table between players, signaling its horizontal translational motion. Proper recognition of this bouncing-sound is causal for interpretations such as *someone is playing ping pong*, but events are typically not verbalized based on their acoustic features (Ballas & Howard, 1987; van Petten & Rheinfelder, 1995: 486). In fact, it would be unintuitive and not sufficiently informative (in German) to describe what is happening in this stimulus with, e.g., *a ping pong ball is bouncing across a table*. Nonetheless, the recognition of item 29 and its

adequate verbalization depends on listeners' knowledge of how ping pong typically sounds when it is played and the schema-based inference about what causes this typical sound (e.g., bat-wielding players positioned at the short ends of a specific type of table). The auditory input they perceive is caused by a series of motion events, and it is these events that they will need to have integrated successfully into a macro-event of *ping pong playing* before they are able to retrieve the lexemes necessary for *someone is playing ping pong*. The process by which participants arrive at such homogenous interpretations on a macro-level (which is beneficial for efficiently responding to the task question *What is happening?*) from auditory stimuli on the micro-level (motion event of the bouncing ball) requires that they successfully compose the auditory object from the auditory events (cf. *Auditory Scene Analysis*, Chapter 2.2). Even though their verbal response may not foreground the motion event, the construal underlying the response necessitated that the motion event was integrated into a more unified, more easily labeled, and articulable event model (see Ballas, 1993: 254; Vanderveer, 1979; Lemaitre et al., 2010 for similar procedure) — especially since they produce a concrete utterance under time pressure.

A similar argument holds for one-word responses. Although rare (6% of all verbal data), one-word responses were mainly nominalized verbs (e.g., *langsames Tropfen*, slow dripping), present participles of verbs (*eine fallende Geldmünze*, a falling coin) or proper nouns (e.g., *Regen*, rain), all of which labeled the macro-event or singled out the audible entity. The scarcity of verbal information makes such responses difficult to classify as motion events. For instance, item 32 replayed the sirens of a firetruck driving in a city environment, which was often described nominally as *fire engine sirens*. The salience of the sirens in the audio stimulus may cause the event to be conceived of as an auditory icon[68], without necessarily considering that it is usually caused by large vehicles in translational motion. On the other hand, item 11 played the hissing sound of a skyrocket that exploded in the distance shortly after. This hissing sound is what disambiguates the subsequent explosion sound, making it clear that this is an ascending and exploding skyrocket rather than, e.g., a popping balloon. Consequently, this hissing sound (auditory event) is fundamental for recognizing the macro-event *fireworks* (auditory object), as responded by one-third of all participants, it is likely that the micro-event played a crucial role in macro-event construal. However,

[68] The same way that a specific ringtone of a cellphone indicates an incoming text message versus a phone call.

since a one-word utterance does not imply that the flight trajectory of the skyrocket was conceptualized in the same manner as for overt verbalization, it remains uncertain whether participants construed such events as having vertical directionality — as Talmy (2000b) would claim. In fact, this is part of what this project examines in spontaneous non-visual gaze behavior.

The meticulous approach presented in this section was necessary due to variations in item verbalization. Participants described the stimuli freely and naturally, not instructed to focus on the motion per se. Therefore, stimuli were sometimes described with less detail and often without motion verbs in the predicate, despite the underlying event existing only because of an entity in motion. Thus, relying solely on the linguistic output is not a fruitful method for examining how participants conceptualized the translational motion events here, and doing so would have excluded numerous trials from analysis.

While individual descriptions varied, all participants were exposed to the same auditory input. This is why the primary criterion for classification was that the response confirmed that the participant had correctly identified the integral motion event in the environmental sound, even if the verbal output did not explicitly emphasize movement. The underlying assumption is that listeners first perceive the motion event and then integrate it into a unified event model, which may not principally activate lexemes associated with movement. This tendency was likely influenced by verbalization demands under time pressure, prompting participants to select the most immediately accessible label. Nevertheless, at its core, the perceived event remains one of motion (Talmy, 2000b).

### 4.5.5. Analysis procedure

One central objective of this study is to compare eye movements during cognitive processing of motion events in language production and language comprehension. This demands similar analysis procedures in both experiments. Like in Experiment 1, the dependent variables were travel distance(x), travel distance(y), and saccade rate. Independent variables included *motion event interpretation* (3 levels), which was determined for each trial based on the subject's verbal output (see above). Similarly, predictors included subjects' introspective ratings of *visualization intensity* (ordinal factors) and *epoch duration* (standardized continuous numeric). Task phases

(encoding vs. recall) do not apply in the design of Experiment 2 but a distinction of *trial epochs* (4 levels: audio replay, pre-button, pre-voice, articulation) may serve as a rough analogue given that they are also related to different mental operations in task-solving.

Predictor variables were again dummy coded for neat model summary interpretation, with reference levels set to non-motion interpretations, epochs in relation to audio replay, and a self-reported visualization intensity of 5 (see Table 4-7 above). Controlling for inter-subject variance, a grouping term for subject was included with a random slope for saccade rate. Specific hypotheses required modifications of this syntax or an exclusion of levels in categorical predictors through subsetting, which will be discussed in the corresponding sections. In sum, despite minor modifications, final model syntax remains comparable to that of Experiment 1:

| | |
|---|---|
| *travel distance(x,y)* | *~ motion event interpretation + visualization + trial epoch + epoch duration* + (1+*rate*\|*subject*) |
| *saccade rate* | *~ motion event interpretation + visualization + trial epoch + epoch duration* + (1\|*subject*) |

### 4.5.6. Formal hypotheses

The goal of Experiment 2 was to examine whether oculomotor measures are systematically affected during language production when participants are tasked with spoken description of motion events (critical condition) presented as auditory non-verbal stimuli. The following hypotheses were formulated:

H4) Hypothesis 4: Motion event interpretations increase travel distance in the axis parallel to the event's inherent direction.

$H_1$: $\text{Travel distance(x)}_{\text{HORIZONTAL}} > \text{Travel distance(x)}_{\text{VERTICAL AND NO-MOTION}}$

$\text{Travel distance(y)}_{\text{VERTICAL}} > \text{Travel distance(y)}_{\text{HORIZONTAL AND NO-MOTION}}$

$H_0$: $\text{Travel distance(x)}_{\text{HORIZONTAL}} \leq \text{Travel distance(x)}_{\text{VERTICAL AND NO-MOTION}}$

$\text{Travel distance(y)}_{\text{VERTICAL}} \leq \text{Travel distance(y)}_{\text{HORIZONTAL AND NO-MOTION}}$

H5) Exploratory Hypothesis 5: Oculomotor activity in pre-articulation speech planning stages is different from that during stimulus perception.

$H_1$: $\text{Travel distance \& rate}_{\text{PRE-ARTICULATION}} \neq \text{Travel distance \& rate}_{\text{STIMULUS REPLAY}}$

$H_0$: $\text{Travel distance \& rate}_{\text{PRE-ARTICULATION}} = \text{Travel distance \& rate}_{\text{STIMULUS REPLAY}}$

H6) Exploratory Hypothesis 6: Stronger effects on oculomotor activity are observed when utterances explicitly reference spatial components of motion events compared to when such references remain implicit.

$H_1$: Travel distance & $rate_{EXPLICIT}$ > Travel distance & $rate_{IMPLICIT}$

$H_0$: Travel distance & $rate_{EXPLICIT}$ ≤ Travel distance & $rate_{IMPLICIT}$

### 4.5.7. Results

#### 4.5.7.1. Hypothesis 4

Essentially, Hypothesis 4 tested whether travel distance(x/y) would increase more strongly when the verbal response indicated a motion event interpretation than when it did not. The analysis is motivated by the notion that simulations are employed in message generation and affect saccadic travel distances throughout the language production process.

| | **Travel distance(x)** | | | **Travel distance(y)** | | | **Saccade rate** | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Estimate* | *CI* | *p* | *Estimate* | *CI* | *p* | *Estimate* | *CI* | *p* |
| (Intercept) | 1.82 | 1.47 – 2.16 | <0.001 | 2.48 | 2.18 – 2.77 | <0.001 | -0.17 | -0.30 – -0.04 | 0.01 |
| Motion interpretation [vertical] | 0.04 | -0.01 – 0.09 | 0.12 | 0.02 | -0.04 – 0.08 | 0.45 | 0.02 | -0.01 – 0.05 | 0.18 |
| Motion interpretation [horizontal] | 0.02 | -0.03 – 0.06 | 0.53 | -0.01 | -0.07 – 0.04 | 0.68 | 0.03 | 0.00 – 0.06 | 0.02 |
| Visualization intensity [4] | 0.12 | -0.30 – 0.53 | 0.59 | 0.03 | -0.31 – 0.37 | 0.87 | -0.02 | -0.20 – 0.17 | 0.86 |
| Visualization intensity [3] | -0.37 | -0.84 – 0.11 | 0.13 | -0.12 | -0.50 – 0.27 | 0.56 | 0.02 | -0.19 – 0.23 | 0.85 |
| Visualization intensity [2] | 0.18 | -0.54 – 0.89 | 0.62 | 0.33 | -0.25 – 0.91 | 0.27 | -0.02 | -0.34 – 0.30 | 0.90 |
| Epoch duration (z-scored) | 1.91 | 1.81 – 2.01 | <0.001 | 1.80 | 1.68 – 1.92 | <0.001 | -0.44 | -0.49 – -0.39 | <0.001 |
| Epoch [pre-button] | 0.09 | 0.01 – 0.17 | 0.02 | 0.21 | 0.12 – 0.30 | <0.001 | 0.18 | 0.13 – 0.22 | <0.001 |
| Epoch [pre-voice] | 0.17 | 0.10 – 0.25 | <0.001 | 0.47 | 0.38 – 0.56 | <0.001 | 0.15 | 0.11 – 0.19 | <0.001 |
| Epoch [articulation] | -1.92 | -2.10 – -1.75 | <0.001 | -1.63 | -1.84 – -1.42 | <0.001 | 0.72 | 0.62 – 0.82 | <0.001 |

***Table 4-11**:* Linear mixed model results for Hypothesis 4. Estimates, 95% confidence intervals (*CI*), and p-values are reported for three dependent variables: travel distance in the x-axis, travel distance in the y-axis, and saccade rate. Predictors include interpreted movement direction, visualization intensity, epoch duration (z-scored), as well as the different epochs. The reference levels for categorical predictors are movement direction [no motion], visualization intensity [5], and epoch [audio replay]. Random effects and model fit are listed in Appendix B5.

| | **Travel distance(x)** | | **Travel distance(y)** | | **Saccade rate** | |
|---|---|---|---|---|---|---|
| | Estimate (*df*) | *p* | Estimate (*df*) | *p* | Estimate (*df*) | *p* |
| Motion interpretation | $\chi^2(2) = 2.5$ | 0.28 | $\chi^2(2) = 1.2$ | 0.55 | $\chi^2(2) = 5.3$ | 0.07 |
| Visualization intensity | $\chi^2(3) = 4.3$ | 0.23 | $\chi^2(3) = 2$ | 0.57 | $\chi^2(3) = 0.1$ | 0.99 |
| Epoch | $\chi^2(3) = 509$ | < 0.001 | $\chi^2(3) = 245$ | < 0.001 | $\chi^2(3) = 572$ | < 0.001 |
| Epoch dur. (z-scored) | M = 1.9 | < 0.001 | M = 1.8 | < 0.001 | M = -0.4 | < 0.001 |

***Table 4-12**:* Wald-test results show significant associations of predictors with response variables in Experiment 1. For categorical predictors, $\chi^2$ (chi square) represents the significance of categorical differences. For continuous predictors, M represents the mean estimated effect size on the response variable.

Statistical modeling revealed that the overall factor of expressed movement direction (vertical, horizontal, no-motion) was not associated with changes in travel distance(x) ($\chi^2(2) = 2.5$, $p = 0.28$). On the distinct variable levels, the influence of vertical motion event descriptions was numerically slightly stronger (Est. = 0.04, $p = 0.12$) than that of horizontal motion event descriptions (M = 0.02, $p = 0.53$), although neither reached the 0.05 significance threshold. Participants' self-reported strength of visualization did not affect this measure ($\chi^2(3) = 4.3$, $p = 0.23$). The strongest statistical association with changes in horizontal travel distance had the variable epoch ($\chi^2(3) = 509$, $p < 0.001$), suggesting distinct effects of cognitive processing associated with distinct language production stages. Epoch duration (Est. = 1.9, $p < 0.001$) is strongly positively correlated with travel distance(x).

When vertical travel distance was modelled as the dependent variable, the strongest predictors were, again, epoch ($\chi^2(3) = 245$, $p < 0.001$) and epoch duration (Est. = 1.8, $p < 0.001$). Stimulus interpretations overall did not induce significant changes in vertical travel distance ($\chi^2(2) = 1.2$, $p = 0.55$) and neither did participants' experienced visualization strength ($\chi^2(3) = 2$, $p = 0.57$), suggesting that vertical travel distance was not influenced by these experimental conditions.

In sum, results speak for rejection of Hypothesis 4. Statistical associations between participants' interpretations of motion events, where the figure moves in horizontal or vertical direction, and the direction of their eye movements were not detected in the full models for language production.

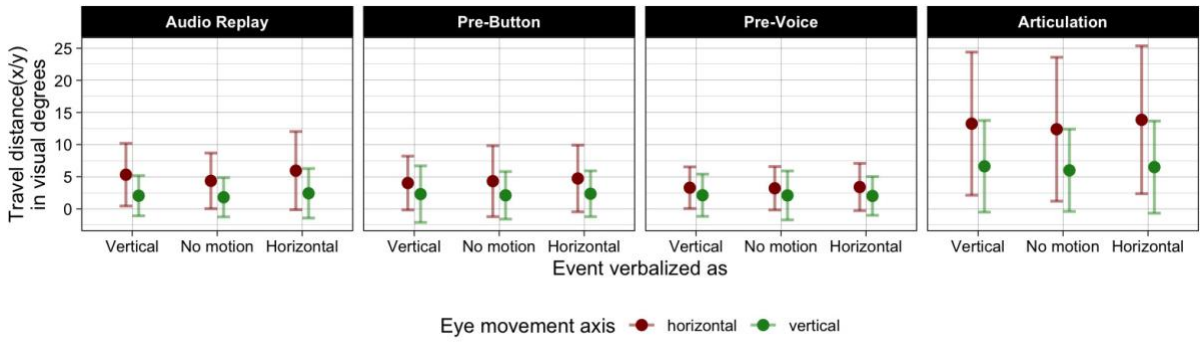


***Figure 4-13**:* Travel distances (horizontal = red) and (vertical = green) in visual degrees (y-axis) by movement direction conditions (x-axis) and split by epoch (columns). Error bars show 1 SD. The 'outer' epochs, audio replay and articulation show larger mean travel distances and variance than the 'inner' epochs. Travel distances of audio replay, in the leftmost panel, hint at differences rooted in the direction implied by the motion event interpretation.

***Interim discussion: Language production captures attentional resources***

The full models were run on combined data of all trial epochs to see if participants' interpretations had overall effects on eye movement behavior during language production. The model specification purposefully blindfolds statistical analysis to the qualitative differences between the processes active throughout each epoch (e.g., perception, categorization, subprocesses of language production). In other words, effects were modeled as if there was a unique and homogenous cognitive process upon which simulations from motion event interpretations may have an impact. Visual inspection of the travel distances(x/y) in Figure 4-13 above suggests that there may be differences between the epochs, and this assumption is statistically supported (cf. Table 4-11): the factor epoch was strongly associated with differences in horizontal travel distances in the above models. The pre-button (Est. = 0.09, $p = 0.02$) and the pre-voice (Est. = 0.17, $p < 0.001$) epochs saw an increase, and travel distance(x) eventually decreased in the articulation epoch (Est. = -1.9, $p < 0.001$).

Previous research proposed that the mental operations associated with these epochs may require shifts of attentional focus between different representational media (e.g., *language* and *situated simulation*; Barsalou et al., 2008), which may further co-occur with systematic changes in oculomotor activity (Ehrlichman & Barrett, 1983; Ehrlichman & Micic, 2012). For instance, when participants are busy with online auditory processing (during audio replay) or speech output monitoring (during articulation), they are more focused on information that is external (bottom-up) (Radvansky & Zacks, 2014), whereas during lexical retrieval, syntactic parsing and phonological encoding, attentional resources are required internally for fast and efficient production of linguistic utterances (cf. Ehrlichman & Barrett, 1983; Micic et al., 2010). Spontaneous eye movements may exhibit systematically variant characteristics in these different trial epochs, and it is theoretically relevant to examine this variation (Barsalou et al., 2008).

Consequently, the data were split into two different subsets and model syntax for the subset data was kept identical to the full data models. The first subset compiled the audio replay and articulation epochs (exteroceptive / outward-directed attention: stimulus perception and articulation monitoring) and the second consisted of the pre-button and pre-voice epochs (interoceptive / inward-directed attention). Note that this is not a claim of a one-to-one correspondence between these epochs and the temporal unfolding of mental activity. Undoubtedly, incremental processing preempts such

claims (see Ch. 2.1 and 5.3.1.2). The epoch distinction is a rough approximation to processing steps that distinctly mobilize global attentional resources.

Subset modelling relativizes the above findings (cf. Appendix B6 for full results tables). Participants' interpretations are not a significant predictor for travel distance(x) in the pre-button and pre-voice epochs ($\chi^2(2) = 2.1$, $p = 0.36$), but become more significant in the audio replay and articulation epochs ($\chi^2(2) = 10.5$, $p = 0.005$). Coefficient estimates for horizontal interpretations (Est. = 0.04, $p = 0.13$) were approaching significance and slightly lower than those for vertical interpretations (Est. = 0.08, $p = 0.001$). Modeling of travel distance(y) did not yield significant results and is not reported.

To summarize, effects of motion event interpretation on travel distance(x) do not hold across all epochs equally. Subset model results confirm this effect for trial epochs in which attentional resources seem to be weighed in favor of bottom-up information streams but not for epochs in which resources are bound by online processes of speech production. Prematurely concluded, no evidence was found that eye movements were systematically affected by stimulus conditions while the cognitive system is focused on speech planning. This finding warrants further discussion (see Chapter 5.2.7) and justifies exploration in Hypothesis 5 below.

#### 4.5.7.2. Exploratory Hypothesis 5

Based on the findings in Hypothesis 4, Exploratory Hypothesis 5 states that travel distances and saccadic rate are modulated by trial epoch. Specifically, it examines whether effortful internal processes of speech planning affect oculomotor activity due to attentional resources being drawn away from bottom-up sensory processing to reduce cognitive load and prevent interference with subprocesses of language production (Ehrlichman & Micic, 2012; Smallwood & Schooler, 2006).

***Preprocessing: z-scoring of epoch durations***

The coefficients of the full model for Hypothesis 4 (cf. Table 4-11) attested that, in comparison to audio replay, travel distance in both axes increased in the pre-button (Est. = 0.09, p = 0.02) and pre-voice epochs (Est. = 0.17, p < 0.001), and eventually decreased in the articulation epoch (Est. = -1.9, p < 0.001). Presumably, during pre-

button and pre-voice, eye movements covered more average travel distance. This seems odd because the mean durations of these epochs (see Table 4-10 above) are significantly shorter ($F(2) = 1510$, $p < 0.001$)[69]. On second thought, it is mathematically plausible that a single saccade during a short epoch accumulated a larger travel distance relative to epoch duration than many smaller saccades in a longer epoch. To circumvent this dependence on epoch duration, durations of each language production stage were standardized separately (z-scored), which eliminated quantitative differences in duration means.

In addition to z-scoring, the data from the articulation epoch was excluded from this analysis. The articulation epoch had the same duration in every trial and, thus, exhibits no variance. Even if its durations were z-scored, in statistical terms, the articulation epoch would not qualify as a continuous predictor but rather a categorical one, whose influence is not comparable, neither conceptually nor parametrically, to that of the other epochs. Additionally, given individual variability in task efficiency, subjects' cognitive occupation with the verbal response may end abruptly upon their impression that a good-enough response was given. Awaiting the next trial, they may idle for the larger part of the articulation epoch instead of attending to the verbalized event, yet all the while their oculomotor activity is measured and analyzed as a signal of online processing.[70]

A final modification to model structure concerns the seemingly counterintuitive exclusion of visualization as a predictor. After all, the logic of Hypothesis 5 requires that the influence on oculomotor activity of internal processes other than language production, such as the experience of emerging visual imagery, be controlled for. However, neither model summaries of Hypothesis 4 nor model comparisons regarding Hypothesis 5 yielded significant interactions of visualization intensity with travel distance, let alone improvement of model fit. In other words, the exclusion of visualization intensity as a predictor does not make a difference from a statistical perspective. In sum, model syntax for Hypothesis 5 was amended the following way:

[69] The pre-button epoch is 842 ms shorter than the audio replay epoch, and pre-voice is 1169 ms shorter; in turn, pre-voice is also 327 ms shorter than the pre-button epoch.

[70] Additionally, the motions of participants' facial muscles during speech may affect eye-tracking accuracy due to squinting and slight head movement.

| | |
|---|---|
| *travel distance(x/y)* | *~ trial epoch + epoch duration (z-scored) + motion event interpretation* + (1+*rate*\|*subject*) |
| *saccade rate* | *~ trial epoch + epoch duration (z-scored) + motion event interpretation* + (1\|*subject*) |

### *Results*

| | Travel distance(x) | | | Travel distance(y) | | | Saccade rate | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Estimate* | *CI* | *p* | *Estimate* | *CI* | *p* | *Estimate* | *CI* | *p* |
| (Intercept) | 1.44 | 1.25 – 1.63 | <0.001 | 0.10 | -0.06 – 0.26 | 0.22 | 0.06 | -0.01 – 0.13 | 0.10 |
| Epoch [pre-button] | -0.51 | -0.58 – -0.44 | <0.001 | -0.37 | -0.46 – -0.29 | <0.001 | 0.33 | 0.30 – 0.37 | <0.001 |
| Epoch [pre-voice] | -0.72 | -0.78 – -0.66 | <0.001 | -0.38 | -0.45 – -0.31 | <0.001 | 0.38 | 0.34 – 0.41 | <0.001 |
| Epoch duration (z-scored) | 0.49 | 0.46 – 0.51 | <0.001 | 0.45 | 0.42 – 0.49 | <0.001 | -0.13 | -0.14 – -0.11 | <0.001 |
| Motion interpretation [vertical] | -0.01 | -0.08 – 0.05 | 0.71 | -0.01 | -0.08 – 0.07 | 0.9 | 0.01 | -0.03 – 0.05 | 0.56 |
| Motion interpretation [horizontal] | -0.04 | -0.11 – 0.02 | 0.18 | -0.03 | -0.10 – 0.05 | 0.48 | 0.01 | -0.03 – 0.05 | 0.57 |

***Table 4-14**:* Linear mixed model results for the epoch comparison of Hypothesis 5. Estimates, 95% confidence intervals (*CI*), and p-values are reported for three dependent variables: travel distance in the x-axis, travel distance in the y-axis, and saccade rate. Predictors include the different epochs, epoch duration (z-scored), as well as interpreted movement direction. The reference levels for categorical predictors are movement direction [no motion], and epoch [audio replay]. Random effects and model fit are provided in Appendix B7.

| | Travel distance(x) | | Travel distance(y) | | Saccade rate | |
|---|---|---|---|---|---|---|
| | Estimate (*df*) | *p* | Estimate (*df*) | *p* | Estimate (*df*) | *p* |
| Epoch | $\chi^2(2) = 560$ | < 0.001 | $\chi^2(2) = 128$ | < 0.001 | $\chi^2(2) = 586$ | < 0.001 |
| Motion interpretation | $\chi^2(2) = 1.88$ | 0.39 | $\chi^2(2) = 0.53$ | 0.77 | $\chi^2(2) = 0.46$ | 0.79 |

***Table 4-15**:* Wald-test results for the epoch comparison (H5).

Model output indicates that, on average, travel distance in both axes is shorter in the intermediate epochs as opposed to the audio replay epoch. The analysis of horizontal travel distance yields strong effects for epoch duration (Est. = 0.49, $p < 0.001$) and for epoch ($\chi^2(2) = 560$, $p < 0.001$), with travel distance(x) significantly reduced in pre-button (Est. = -0.51, $p < 0.001$) and pre-voice (Est. = -0.72, $p < 0.001$) as opposed to the reference epoch audio replay. A similar reduction holds for vertical travel distance (pre-button (Est. = -0.37, $p < 0.001$) and pre-voice (Est. = -0.38, $p < 0.001$), though overall associations were weaker (epoch: $\chi^2(2) = 128$, $p < 0.001$; epoch duration: Est. = 0.45, $p < 0.001$). Consequently, when epoch durations are equal, participants' saccades covered shorter travel distances in epochs dedicated to online processes of language production. Reinforcing the results of Hypothesis 4, neither model detected systematic changes in travel distances to be associated with participants' motion event

interpretation, neither in the horizontal ($\chi^2(2) = 1.88$, $p = 0.39$) nor in the vertical axis ($\chi^2(2) = 0.53$, $p = 0.77$).

Like before, saccade rate was modelled as the dependent variable to complement the travel distance data. Identical fixed-effects structure and data sets were used. Central effects of epoch ($\chi^2(2) = 586$, $p < 0.001$) confirm that saccade rate increased significantly in both trial epochs pre-button (Est. = 0.33, $p < 0.001$) and pre-voice (Est. = 0.38, $p < 0.001$) with audio replay as reference. The increase in rate is surprising in view of the above finding that travel distances were reduced. In other words, during linguistic encoding, participants initiated more saccades but accumulated smaller travel distances.

To assess this statistically, an additional mixed model was fit to examine the interaction between saccade rate and epoch in their effects on horizontal travel distance(x) post-hoc (cf. Appendix B8 for the results table). The model confirmed the expected interaction: while saccade rate (Est. = 1.25, $p < 0.001$) and both epochs (pre-button: Est. = 0.99, $p < 0.001$; pre-voice: Est. = 0.23, $p < 0.001$) were each associated with significant increases in travel distance (x), their interaction term revealed a reduction in travel distance (x) at higher saccade rates in both epochs (pre-button: Est. = -1.06, $p < 0.001$; pre-voice: Est. = -0.75, $p < 0.001$). This suggests that the effect of saccade rate on travel distance (x) was reduced in comparison to the audio replay epoch. In other words, while participants produced more saccades in the immediate language production epochs, the horizontal travel distance covered per saccade was smaller.

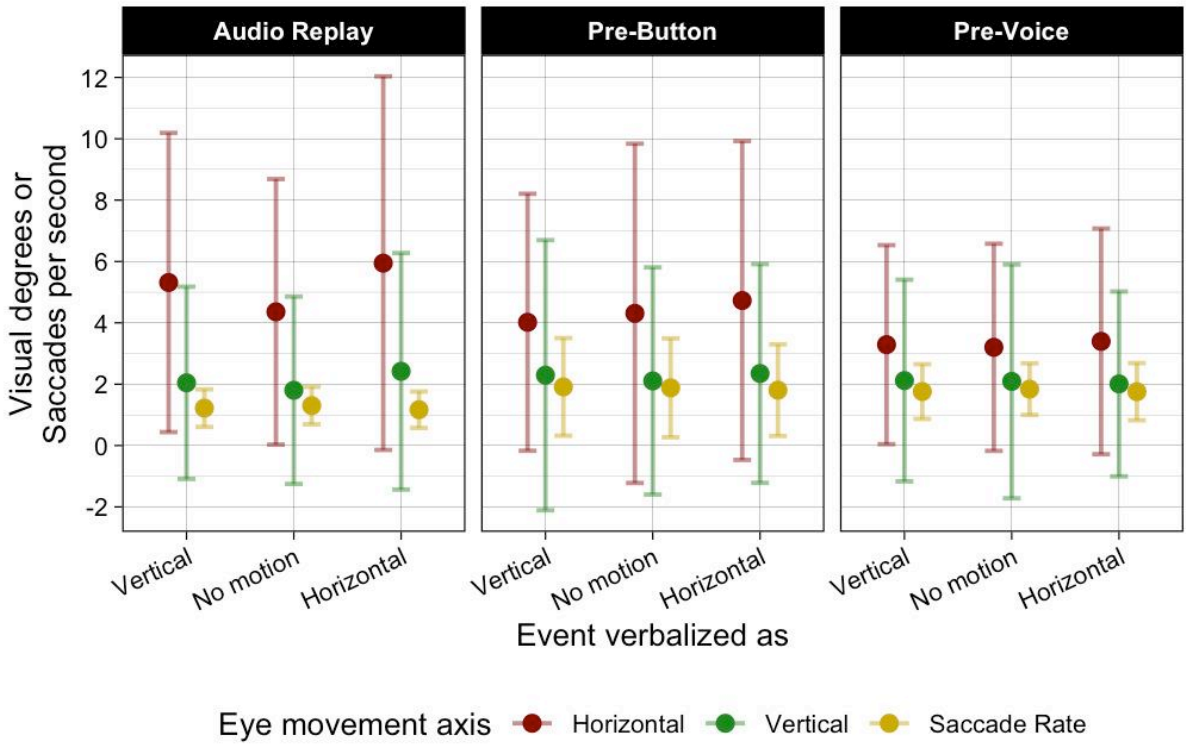


***Figure 4-16**:* Travel distances (horizontal = red) and (vertical = green) in visual degrees, as well as saccade rate (yellow) in saccades per second, by interpreted movement direction condition (x-axis) in different epochs (columns). Error bars show 1 SD. Saccade rates (yellow) significantly increase in the intermediate stages (middle and right panel) as compared to audio replay (left panel). However, this increase in rate does not bring about similarly strong increases in travel distance in the intermediate epochs — while the yellow dots rise, the red and green dots sink.

### *Interim discussion*

It is difficult to judge whether this finding may be an artefact of data transformation, considering that raw rate and travel distance distributions exhibit strong right skews and are both on different scales. Transforming the raw variables to achieve normal distributions may have obscured conditional differences. Yet, both raw variables were derived from the same signal (*saccadR* output). Therefore, they differ only in how they represent the measured saccades – either in terms of temporal frequency or spatial extension. Any distributional characteristics of the raw variables related to experimental conditions would be preserved after data transformation.

To sum up, Hypothesis 5 investigated whether speech planning would influence oculomotor activity differently. Analyses suggest smaller spatial extension of eye movements, as if participants were staying in a smaller area with their eyes and keeping their eyes rather still.

#### 4.5.7.3. Exploratory Hypothesis 6

The previous results attest movement direction effects on travel distance(x) for the audio replay and articulation phases (Hypothesis 4), but these effects did not hold in the speech planning epochs, making it seem that oculomotor activity is different as a factor of epoch (Exploratory Hypothesis 5). One question that remains unanswered in these analyses is whether movement direction effects in speech planning epochs may have occurred only then, when movement direction was explicitly verbalized.

Despite the classification of utterances as interpretations of motion events (cf. Ch. 4.5.4), it remains uncertain whether participants conceptualized movement direction or other spatial components when they did not overtly realize these components in their responses. Since one of this study's central assumptions is that they do, regardless of the semantic structure of their utterance (as is claimed by, e.g., Talmy, 2000b; Papafragou et al., 2008; Trueswell & Papafragou, 2010), such incomplete motion event descriptions must not be excluded; instead, they must be the control condition for Exploratory Hypothesis 6. This hypothesis departs from the premise that a linguistic utterance is a measurable, surface-level expression of an underlying mental representation (i.e., a construal). In other words, what we conceive as salient in a mentally represented event is assumed to be revealed through the way we express this event in language (Slobin, 1996; Gerwien & von Stutterheim, 2018; Gerwien & von Stutterheim, 2022). This notion presumes that conceptualization is verbal to some degree, with lexemes and morphosyntactic structure of the utterance reflecting the speaker's conceptualization (cf. findings on visual attention by von Stutterheim et al., 2012; Flecken et al., 2015).

Exploratory Hypothesis 6 predicts that explicit linguistic cues to movement direction or goal (e.g., 'upwards' or 'to the floor') will amplify eye movement effects, yielding larger travel distances along the corresponding axes. If speakers simulated an idealized directional path during message generation, making it salient enough to surface in the response utterance, then eye movements should reflect that simulation in the early phases of speech planning. Thus, utterances containing overt directional expressions provide a critical test of how simulations of movement direction can influence oculomotor behavior.

Statistical analysis related to Exploratory Hypothesis 6 hinges on the comparison of two subsets of the eye-tracking data. First, from participant responses

with explicit expression of a conceptual component of space (critical condition), that is, all cases where directional or endpoint information was overtly stated in the utterance. Second, implicit cases where a motion event was understood but the directional component was not explicitly verbalized (control condition). If explicit cases showed significantly longer travel distances(x/y), this would indicate that a more visuo-spatially detailed simulation underpinned early processes of language production (i.e., message generation) and resulted in the activation of explicit lexemes. If no significant effect was found, it would indicate that explicit verbalizations of directional information are not primarily driven by simulations of movement, contrary to what is expected from simulation-based accounts of conceptual representation.

In short, Exploratory Hypothesis 6 rests on the idea that, if simulation is a fundamental mechanism of conceptual representation, then message generation and construal for verbalization rely on simulation, too. In logical terms, this is based on reverse inference from the speech output to the underlying conceptual representation. If participants verbalized these spatial components, their assumed conceptual salience during message generation persisted through all subsequent processes until explicit articulation. Oculomotor activity related to these salient components may indicate here whether simulation is employed as a means of construal.

***Preprocessing and model syntax***

To investigate this, all verbal responses that were classified as revealing a motion event interpretation were labeled as explicitly referring to GROUND/GOAL, DIRECTION, both or neither (see Chapter 4.5.4). This was included in the model as a factor with 4 different levels, termed *explicit constituent*. The reference level was 'neither', that is, implicit motion event interpretations. Non-motion interpretations were discarded to isolate how these explicit verbalizations potentially boost effects in comparison to motion event verbalizations without them. Theoretically, such effects on oculomotor activity occur especially prior to articulation. Hence, as in Exploratory Hypothesis 5, data from the articulation epoch was excluded and duration differences between epochs were nivellated by z-scoring. The dataset consisted of 1214 trials, with 677 responses categorized as horizontal and 537 as vertical. On average, each participant contributed 29 ± 5 trials. Of the utterances produced, an average of 10 ± 4 per participant left direction implicit (total: 849), while 3 ± 2 explicitly referred to direction (total: 188), and

another 3 ± 2 included references to both ground and direction (total: 174). To perform linear regression, model syntax required the following structure:

| | |
|---|---|
| *travel distance(x,y)* | *~ explicit constituent + motion interpretation + visualization + trial epoch + epoch duration (z-scored)* + (1+*rate*\|*subject*) |
| *saccade rate* | *~ explicit constituent + motion interpretation + visualization + trial epoch + epoch duration (z-scored)* + (1\|*subject*) |

### *Results*

| | Travel distance(x) | | | Travel distance(y) | | | Saccade rate | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Estimate* | *CI* | *p* | *Estimate* | *CI* | *p* | *Estimate* | *CI* | *p* |
| (Intercept) | 1.38 | 1.08 – 1.67 | <0.001 | -0.05 | -0.32 – 0.22 | 0.72 | 0.01 | -0.13 – 0.16 | 0.88 |
| Explicit constituent [ground] | -0.26 | -0.92 – 0.39 | 0.43 | 0.26 | -0.53 – 1.05 | 0.52 | -0.14 | -0.53 – 0.25 | 0.48 |
| Explicit constituent [direction] | -0.01 | -0.10 – 0.09 | 0.90 | -0.00 | -0.12 – 0.11 | 0.94 | 0.00 | -0.06 – 0.06 | 0.96 |
| Explicit constituent [ground & direction] | 0.07 | -0.03 – 0.17 | 0.15 | -0.07 | -0.18 – 0.05 | 0.25 | 0.03 | -0.03 – 0.09 | 0.31 |
| Motion interpretation [vertical] | 0.02 | -0.05 – 0.09 | 0.62 | 0.02 | -0.06 – 0.11 | 0.56 | -0.00 | -0.05 – 0.04 | 0.86 |
| Visualization intensity [2] | 0.17 | -0.25 – 0.59 | 0.43 | 0.04 | -0.35 – 0.42 | 0.86 | 0.01 | -0.18 – 0.20 | 0.90 |
| Visualization intensity [3] | -0.34 | -0.82 – 0.14 | 0.17 | -0.06 | -0.50 – 0.38 | 0.78 | 0.08 | -0.13 – 0.29 | 0.47 |
| Visualization intensity [4] | 0.18 | -0.54 – 0.90 | 0.63 | 0.41 | -0.25 – 1.07 | 0.22 | 0.07 | -0.23 – 0.38 | 0.65 |
| Epoch [pre-button] | -0.49 | -0.58 – -0.40 | <0.001 | -0.34 | -0.45 – -0.23 | <0.001 | 0.35 | 0.29 – 0.40 | <0.001 |
| Epoch [pre-voice] | -0.70 | -0.78 – -0.62 | <0.001 | -0.33 | -0.43 – -0.23 | <0.001 | 0.35 | 0.30 – 0.40 | <0.001 |
| Epoch duration (z-scored) | 0.47 | 0.43 – 0.51 | <0.001 | 0.43 | 0.39 – 0.48 | <0.001 | -0.11 | -0.14 – -0.09 | <0.001 |

***Table 4-17:*** Linear mixed model results for the epoch comparison of Hypothesis 6. Estimates, 95% confidence intervals (*CI*), and p-values are reported for three dependent variables: travel distance in the x-axis, travel distance in the y-axis, and saccade rate. Predictors include the different forms of explicit constituents, interpreted movement direction, visualization intensity, epoch, as well as epoch duration (z-scored). The reference levels for categorical predictors are movement direction [horizontal], epoch [audio replay], visualization intensity [5]. Random effects and model fit are provided in Appendix B9.

| | Travel distance(x) | | Travel distance(y) | | Saccade rate | |
|---|---|---|---|---|---|---|
| | Estimate (*df*) | *p* | Estimate (*df*) | *p* | Estimate (*df*) | *p* |
| Explicit constituent | $\chi^2(3) = 2.9$ | 0.41 | $\chi^2(3) = 1.8$ | 0.61 | $\chi^2(3) = 1.6$ | 0.65 |
| Motion interpretation | $\chi^2(1) = 0.24$ | 0.62 | $\chi^2(1) = 0.34$ | 0.56 | $\chi^2(1) = 0.03$ | 0.86 |
| Visualization intensity | $\chi^2(3) = 4.5$ | 0.21 | $\chi^2(3) = 1.82$ | 0.61 | $\chi^2(3) = 0.71$ | 0.87 |
| Epoch | $\chi^2(2) = 307$ | < 0.001 | $\chi^2(2) = 58$ | < 0.001 | $\chi^2(2) = 247$ | < 0.001 |

***Table 4-18****:* Wald-test results for Hypothesis 6 show associations of collective predictors with response variables. For categorical predictors, $\chi^2$ (chi square) represents the significance of categorical differences.

Statistical modeling revealed no measurable effects of explicit verbalizations on travel distances(x/y) in this dataset (for horizontal travel distance, $\chi^2(3) = 2.9$, $p = 0.41$, whereas for vertical travel distance, $\chi^2(3) = 1.8$, $p = 0.61$). The strongest stochastic relationship with travel distance(x/y) was observed for epoch duration (Est. = 0.47, $p <$

0.001; Est. = 0.43, $p < 0.001$). The strength of this association attests that it explains a large part of the variance, possibly rendering weakly associated variables insignificant.[71] Intensity of visualization had no significant influence ($\chi^2(3) = 4.5$, $p = 0.21$ for x-axis; $\chi^2(3) = 1.82$, $p = 0.61$ for y-axis).

In models on saccadic rate, significant predictors were epoch and epoch duration, which was discussed in a previous section (see Exploratory Hypothesis 5). Explicit verbalizations did not correlate with changes in saccadic rate ($\chi^2(3) = 1.6$, $p = 0.65$). Neither did self-assessed intensity of visualization ($\chi^2(3) = 0.71$, $p = 0.87$).

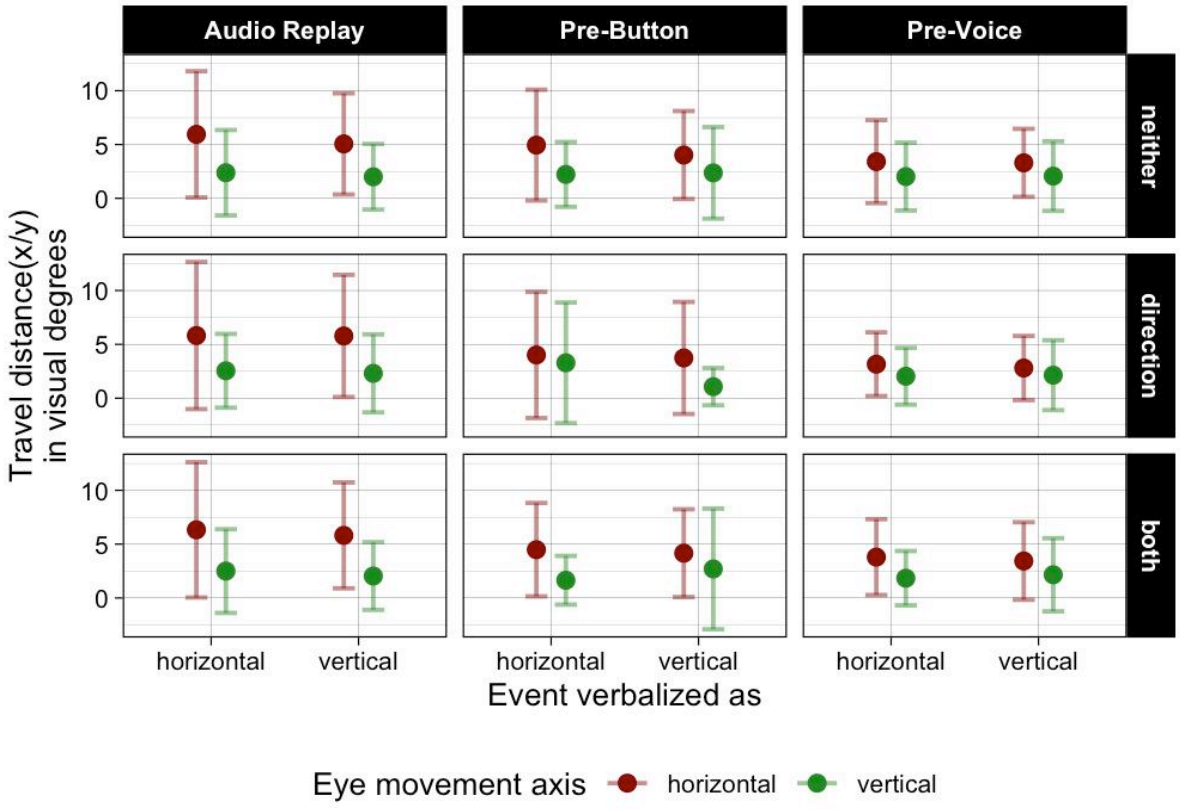


***Figure 4-19**:* Mean travel distances by epoch (columns) and explicit expression of conceptual components of space (rows). Error bars indicate 1 SD. Travel distances of eye movements were not significantly different as a factor of participants' overt verbalizations of spatial components in motion events. Trials with explicit reference to ground (n=4) were not plotted because of small sample size.

### *Interim discussion*

[71] When the predictor *epoch duration* is dropped from model syntax, this model detected a significant increase of horizontal travel distance (x) when the verbal response expressed reference to both GROUND/GOAL and DIRECTION (e.g., *ein Glas fällt auf den Boden runter*), in both horizontal and vertical motion event interpretations (Est. = 0.13, $p = 0.02$) (cf. Appendix B10). However, since models that consider the key control variable *epoch duration* show no such effect, it cannot be interpreted as a robust pattern.

To examine findings of Hypothesis 4 in detail, Exploratory Hypothesis 6 tested whether effects of motion event interpretations on oculomotor activity would be amplified when verbal responses contained overt reference to spatial components of motion. No conclusive evidence was found for simulation-driven eye movement patterns in dedicated language production stages, even when participants verbalized conceptual components of space explicitly. The evidence presented here suggests that the impact of simulation on oculomotor activity during speech planning may be limited, as simulations might not be prioritized by executive attention, but the discussion of the findings is the matter of the next chapter.

### 4.5.8. Summary of the results (Exp. 2)

The objective of Experiment 2 was to examine non-visual eye movements that accompany cognitive processes related to language production. The findings of Experiment 2 thus complement the results of Experiment 1, which investigated language comprehension, the reverse process. Data analysis of Experiment 2 yielded the following results.

Hypothesis 4 tested whether cumulative saccadic travel distance would systematically increase when participants were preparing and uttering a spoken description of a motion event. No significant main effects on travel distance were confirmed, although effects on horizontal travel distance were detected in trial epochs dedicated to stimulus perception and ultimate articulation.

Zooming in on these epochal differences, Exploratory Hypothesis 5 aimed at describing differences in oculomotor activity between epochs requiring attentional resources for perception of the external stimulus as opposed to internal, pre-articulatory processes of message generation and linguistic encoding. Modelling the latter, a noteworthy result emerged as significant: Travel distance, a spatial property, decreased while saccadic rate, a temporal feature of oculomotor activity, increased.

Exploratory Hypothesis 6 investigated whether explicit articulation of spatial motion components amplified potential effects of movement direction on the direction of eye movements prior to speech onset. No such effects were detected, suggesting attenuated influence of simulation on oculomotor activity during language production.

Overall, the analysis of eye movements during Experiment 2 suggests a weaker involvement of simulations during cognitive processes active in language production. Differential influence on measures of oculomotor activity hint at linguistic encoding to

be an effortful process that requires major channeling of attentional resources to subprocesses of language production. To what extent the findings of both experiments converge and how they may substantiate overarching theoretical trends will be discussed in Chapter 5 below.

# 5. Discussion

## 5.1. Summary of the key findings

### 5.1.1. Experiment 1

Experiment 1 aimed to test whether during comprehension of verbal and non-verbal motion event stimuli, spontaneous eye movements reflect that conceptual representation was driven by perceptual simulations.

***Hypothesis 1: Movement direction effects on travel distance in comprehension***

The analyses showed that participants' eye movements were larger when they processed motion events as opposed to events without motion. In both modalities and across task phases, the processing of horizontal motion events correlated with horizontal travel distances larger than those of vertical motion events or non-motion events. The processing of vertical motion events was associated with larger vertical travel distance measures than for non-motion events. However, vertical travel distance was increased even more strongly following horizontal motion events. The motion direction condition was not significantly associated with changes in saccade rate, suggesting that it was larger saccadic amplitude and not a higher saccadic frequency that resulted in the larger travel distances for the respective critical conditions.

These results partially confirm the predictions of Hypothesis 1. Horizontal motion events increase horizontal travel distance, above all else. Though vertical travel distance was enlarged in vertical motion events, horizontal motion events had an even stronger effect on vertical travel distance, contradicting Hypothesis 1 insofar that vertical travel distance was not significantly larger in vertical than in horizontal motion events. Therefore, the initial claim is contradicted that an increase would preferentially influence travel distance in the parallel, rather than the orthogonal, idealized directional axis (Talmy, 2000b; see Chapter 2.6.2). Instead, the findings rather support the claim that the processing of motion, that is, the horizontal and vertical condition combined, as opposed to non-motion events increased travel distances in both axes, suggesting overall larger spatial dispersion of eye movements during motion events.

***Hypothesis 2: Visualization intensity and oculomotor activity***

In addition to the effects reported above, self-reported visualization intensity was another significant predictor of differences in travel distance, particularly for the

horizontal axis and, again, across modalities and task phases. Subjects who reported experiencing vivid visualizations exhibited shorter horizontal travel distance while subjects who self-reported rather weak visualizations made larger eye movements. No such effects were found for vertical travel distance. Interestingly, in the encoding phases, the smaller horizontal travel distance co-occurred with a lower saccade rate in subjects with vivid visualizations, suggesting that their oculomotor activity[72] was attenuated when stimuli were first encountered and encoded into memory. In other words, stronger visualizations appear to coincide with lower frequency and magnitude of eye movements. Strong visualizers seem to have apprehended the input in a different way than weak visualizers.

The predictions of Hypothesis 2 cannot be confirmed. Despite a statistical association between visualization intensity and oculomotor activity, the analysis rejects the claim of embodied cognition that this relationship would be positive, with more oculomotor activity correlated with more intense visualization. The relationship that was found was inverse: the less intense the visualization, the more oculomotor activity was measured.

***Exploratory Hypothesis 3: Task phase differences in oculomotor activity***

Considering that task phase (*encoding* vs. *recall*) had been a highly significant predictor in Experiment 1, an examination of the relationship between task phase and oculomotor activity was warranted. Analysis of Exploratory Hypothesis 3 found that, across variable levels of movement direction, modality, and visualization intensity, travel distances in both axes generally decreased while saccade rate increased in the recall phases when compared to encoding phases. When recalling previously heard stimuli, participants overall executed more eye movements, but the spatial extension of these eye movements was proportionately smaller. In this sense, Exploratory Hypothesis 3 revealed that, while perception of a stimulus and encoding it into memory correlates with lower saccade rates but larger travel distances, proper recognition and recall of stimuli increases rate but shortens travel distances.

An additional finding is that saccade rate was significantly increased in the verbal condition in comparison to the environmental sound condition.

[72] Recall that oculomotor activity was defined as a combined measure of travel distance and rate.

### 5.1.2. Experiment 2

The objective of Experiment 2 was to examine patterns of oculomotor activity for evidence of perceptual simulation of motion event representations constructed for language production.

#### *Hypothesis 4: Partial movement direction effects on travel distance in language production*

While Hypothesis 1 targeted language comprehension, Hypothesis 4 concerned movement direction effects on travel distance during language production. When participants verbalized events with horizontal or vertical movement direction, no significant associations with changes in travel distances resulted in the production task. Hypothesis 4 was therefore rejected.

Post-hoc examinations revealed, however, that movement direction effects on horizontal travel distance were in fact present in the audio replay-epoch — similar to the ones in the comprehension task of Experiment 1. Therefore, Hypothesis 4 was partially confirmed, such that in periods of stimulus perception before speech planning, conceptualization of movement affects horizontal travel distance. Effects on vertical travel distance were not detected.

#### *Exploratory Hypothesis 5: Different oculomotor patterns in speech planning epochs*

Given that the strongest predictor of travel distance in the model testing Hypothesis 4 was trial epoch, Hypothesis 5 explored epochal differences in oculomotor activity. The audio replay epoch served as the baseline against which the immediate speech planning epochs were compared: In the pre-button/pre-voice epochs, travel distance in both axes was significantly decreased, while saccade rate was increased. Although movement direction effects were again absent, this is a noteworthy result. When participants were fully focused on planning and preparing their utterances, that is, after audio offset and before beginning to speak, they make more frequent eye movements, but these eye movements are shorter in length. Consequently, Exploratory Hypothesis 5 revealed different oculomotor activity between epochs, suggestive of shifts in participants' cognitive processing when they turn from stimulus perception to immediate speech planning — this is marked by more frequent but spatially

constrained eye movements, implying that focused processing during linguistic encoding coincides with a defocusing of the conceptual representations previously constructed through perceptual simulations.

***Exploratory Hypothesis 6: Explicit verbalization of spatial motion event components***

Following the logic that overtly verbalized motion event components must have been salient during conceptualization, Hypothesis 6 explored movement direction effects on oculomotor activity in a subset of trials where directional components were verbalized overtly. Guided by the prediction that larger travel distances would be detectable, explicit trials were compared to trials with motion event interpretations where utterances referenced such components of space only implicitly. No significant effects of movement direction on neither travel distance, nor saccade rate were detected. Although directional concepts were activated during conceptualization, the production of such explicit verbalizations did not affect eye movement characteristics in speech planning epochs. Hypothesis 6 was rejected.

## 5.2. Interpretation and discussion of the findings

Analysis results for the main hypotheses are mixed. While the comparison of eye movement patterns between motion events and non-motion events in language and environmental sound comprehension yielded concrete results, the findings for language production are less straightforward.

### 5.2.1. Movement direction: Evidence for simulation in comprehension

For both types of stimuli (verbal vs. environmental sound), specific movement direction effects were found for horizontals, and in a way also for verticals, but vertical travel distance was increased even more strongly for horizontal motion events than for vertical motion events. This suggests that the comprehension of motion events (i.e., verticals and horizontals combined) amplifies spatial dispersion of eye movements, regardless of directional congruence. These findings indicate that perceptual simulation plays a role in constructing conceptual representations during comprehension. They further suggest that event models, as a common representational

structure, capture meaning in both modalities and that movement space was a conceptual component activated in these event models (cf. Ch. 2.4.3).

Contradicting the theoretical assumptions, the travel distance data showed that these simulations were not explicitly driven by implied axial directionality (Spivey & Geng 2001; Spivey, Tyler, Richardson & Young, 2000) or idealized, geometric paths (Talmy, 2000b; Landau, 2017), but by representations of motion extending in space per se. Motion event stimuli contrast with non-motion events in that these controls mainly referenced stationary objects or animate entities that are not required to change their position in space to emit sound (e.g., the honk of a car horn, barking) — they are rather space-less. Accordingly, simulations of stationary objects or non-moving entities do not necessarily activate representations of spatial location changes (cf. Chapter 5.3.3). The result that travel distances were significantly smaller for these control events aligns with this interpretation: eye movements were shorter in control stimuli than in critical motion stimuli because mentally representing the movement of a figure requires representing changes of the figure's position in space (cf. Sima, 2014: 107) — as if the perceptual simulation generated a representation of a figure entity moving along a spatially extended path.

At the same time, eye movements were not more frequent. The statistical models (H1) controlled for the influence of rate on travel distance but the movement direction effects persisted significantly and a separate model confirmed that saccade rates were not different between the conditions. This does not only counter alternative explanations that larger travel distances in the critical conditions could have come about by higher saccade rates. It also supports that participants accumulated more travel distance in the critical conditions with a similar number of saccades as in the control conditions. Therefore, larger travel distances were most likely influenced by condition-specific increases in saccadic amplitudes (similar findings hold for the audio-replay epoch in the production task, cf. H4). Again, the dynamicity of the mental representation underlying comprehension seems to be related to dynamicity of oculomotor behavior. One plausible interpretation of these findings is that simulations of space in motion event representations seem to have impacted travel distances.

### 5.2.2. Stimulus type: Simulation stronger for non-verbal than verbal stimuli

Comparing the eye movement data between the verbal and non-verbal condition yielded a noteworthy result. Travel distances in response to environmental sounds

were larger than those for verbal stimuli, while saccade rates were larger for verbal stimuli than for environmental sounds. Simply put, environmental sound processing co-occurred with fewer but larger eye movements, while verbal stimulus processing triggered more but shorter eye movements. Could this indicate that participants were more inclined to rely on perceptual simulations when the information is coded as experiential as opposed to verbal input?[73]

A few notions are relevant to contemplate this puzzling result. Primarily, the notion of *encoding specificity* (Tulving & Thomson, 1973; Tulving, 1991) concerns empirical evidence that memory recall is improved when stimulus and recall cue are presented in the same modality, or when the retrieval context otherwise matches conditions under which the stimulus was originally encoded. Considering that the task in Experiment 1 explicitly required memory recall, participants might have been more inclined to encode environmental sounds by constructing a compatible perceptual representation using modal simulation (Barsalou, 2008; Kent & Lamberts, 2008) as opposed to memorizing them as words. Environmental sound stimuli are more experiential and may be directly processed and conceptually represented in sensory areas (Körner et al., 2015: 3; Barsalou, 1999). The verbal stimuli, in contrast, presented information not in a sensory-based but symbolic format. Following the principle of encoding specificity again, the verbal stimuli may have prompted participants to memorize linguistic surface forms as opposed to simulated event representations, although it has been shown that surface forms quickly vanish (Kintsch et al., 1990). In other words, the format of the input might influence participants to rely more or less strongly on perceptual simulations.

Still, the linguistic stimuli encode concepts that can be simulated once lexical units are retrieved from the mental lexicon. It might, then, become a matter of processing duration when verbal concepts activate perceptual simulations. Barsalou and colleagues (2008) have proposed that even when processing language, situated simulations occur concurrently to the activation of lexical concepts (Pulvermüller, 1999) and not only after linguistic analysis completed (see Chapter 2.4.1; Kintsch, 1988; Mahon & Caramazza, 2008; Louwerse & Jeuniaux, 2008; cf. Chapter 5.3.5). They acknowledge a gradient between processing of verbal and sensory input: perceptual »experience may activate situated simulations faster than it activates language because simulations are more similar to cue information« (Barsalou et al., 2008: 276).

[73] This relates to issues like amodal language processing, which are discussed below (cf. Ch. 5.2.3).

While verbal stimuli may initially be comprehended using automated processes of language decoding, situated simulations may capture attention afterwards. It is important to note, however, that this is not a claim that language is processed entirely without concurrent simulations or that *language* and *situated simulation* operate in a mutually exclusive fashion. The assumption is that the two modes operate simultaneously but reach peak activation at different latencies after stimulus onset, causing conceptual representations to become more enriched over time, be it through situated simulation or verbal associations.

These assumptions seem to be supported by the differences in oculomotor activity found for the environmental sound and verbal stimuli contrast. Larger travel distances, as a correlate of simulated spatial extension, occur with environmental sound stimuli because they are experiential cues, prompting conceptual representation through a direct activation of perceptual symbols (Barsalou, 1999). In contrast, verbal stimuli initially engage linguistic decoding (Kintsch, 1988), with peak activation of situated simulations emerging only later. Findings from H4 further support this: simulation effects of movement direction were present during audio replay but vanished during speech planning (pre-button/pre-voice epochs), when cognitive processing shifted from perceptual simulation to linguistic encoding. Depending on the specific input being processed, the representational format that gains attentional priority (Liu, 2009; Posner, 1973) biases how conceptual representation unfolds, either towards detailed perceptual simulations or to minimal good-enough representations of verbal utterance meaning (cf. Chapter 5.3.1.3). The shifting between the two proposed modes of conceptual representation is presumably accompanied by qualitative changes of top-down influence on oculomotor behavior.

However, the interpretation that different modes of processing peaked at different timepoints remains speculative due to the time-insensitivity of the dependent variables. Travel distance and rate were aggregated for the whole trial duration and thus do not justify the above claim that simulation-based eye movements occurred only later in verbal trials. If simulation effects on eye movement were delayed until after linguistic decoding (in the verbal modality), one may assume that there simply was not much trial time left for such eye movements to be registered. The aggregated variables can neither confirm nor falsify this hypothetical explanation, despite its plausibility.

Leaving it up to further research to illuminate this train of thought, the present result is that movement direction effects on travel distance were observed as more

strongly for environmental sounds than verbal stimuli. At the very least, this contradicts a radical enactivist conception of cognition: if the enactment of low-level perceptual behavior (here: eye movements) was a necessary condition for high-level cognition (conceptual representation), no differences between movement direction effects on travel distance should have shown between modalities. But, as argued above, the format of the input appears to influence the format of conceptual representation. If simulation was a computational mechanism that was ubiquitous and unrelenting in all cognitive tasks (Barsalou, 2008), differences in eye movements between types of input would be unlikely. Thus, the evidence presented contradicts such strong claims about simulation in cognition.

### 5.2.3. Comprehending language increases saccade rate

While the previous explanation is straightforward for the oculomotor measure of space (travel distance), it does not account for the increases of saccade rates (measure of frequency) in the verbal condition (H1). Why would participants initiate more eye movements when their simulation-related activity is supposedly weaker, less focused, or delayed?

Early studies relating cognitive processing to non-visual oculomotor activity reported similar findings (Antrobus et al., 1964). Ehrlichman and Barrett (1983, Experiment 2), measuring electro-oculography from participants seated in the dark, found that tasks requiring language-based information generation (e.g., »What are four words that mean the same thing as the word funny?«, p. 15) were associated with higher saccade rates than tasks that required participants to generate information using visual imagery (e.g., »What does your stove look like?«, p. 15). In another EOG study, Demarais and Cohen (1998, Experiment 2) found significantly higher saccadic rates when participants were instructed to produce inner speech, either to count or to solve syllogistic problems, as opposed to relying on visual imagery only (see Villena-González et al., 2016, for compatible effects of inner speech).

The present study replicated this trend. Conceptual processing in the verbal modality increased saccadic rate as opposed to the non-verbal condition (H3, H5). Why might this occur? Antrobus and colleagues (1964: 250) assumed that »the rate of eye movement is […] linked to the rate of change of cognitive content«, while Demarais and Cohen (1998: 230) added »visual tasks require less cognitive change than verbal tasks«. Essentially, increases in eye movement rate are interpreted as an indicator of

cognitive load and attention, with high *cognitive change* defined as the conceptual system »“sampling” many different contents, operations, or memory locations« (Ehrlichman & Barrett, 1983: 24). Language processing is characterized by a consistent interplay of encoding or decoding operations that tap into a variety of knowledge bases (grammar, phonology, lexicon, encyclopedic knowledge, context) in a serial-parallel fashion (see Chapter 2.1.1), accumulating cognitive load with the single aim of inferring a message or generating one in utterance production. Even though this serial-parallel processing ultimately converges on a unified outcome, there is no doubt that cognitive processing of language entails lot of cognitive change, increasing the saccadic frequency along the way (Ehrlichman & Barrett, 1983: 23).[74] Ehrlichman and Micic (2012) reformulated this in more general terms of the cognitive architecture of memory, reinterpreting increased eye movement rate as a signal for more effortful long-term memory retrieval processes and a lower saccade rate as indicative of sustained attention to representations currently held in a working memory buffer. Altogether, saccade rates may be influenced by cognitive load, which is increased in language-based tasks.

While this neatly explains the increase of saccade rate in the verbal condition of Experiment 1, it contributes little to understanding the relationship of rate with situated simulation. The discussion below will focus on this link in more detail. With respect to LASS Theory, however, one might prematurely conclude that a frequency increase in eye movement might be a behavioral marker for cognitive activity to peak in the verbal processing mode before it reduces in favor of behavioral expression in the oculomotor dimension of space, with larger travel distances marking a shift to situated simulation.

A pertinent finding to be discussed in this context was obtained in the language production task (Experiment 2). The analysis of H5 confirmed rate differences between the epochs. In the audio replay epoch, travel distances were larger and rates lower in relation to the subsequent pre-button/pre-voice trial epochs, where saccade rate increased while travel distance decreased. This suggests, again, that when participants dedicate their cognitive efforts to the processing of language and prepare

[74] This is debatable from the perspective of Lupyan’s (2012) labeling account, by which the increased saccade rate could actually reflect lower cognitive effort for verbal input. Verbal labels provide direct access to general concepts, letting comprehenders bypass the more extended recognition processes that are necessary for environmental sound comprehension. Verbal labels thus facilitate comprehension and free cognitive resources for active sampling, with simulation subsequently enriching — rather than previously constructing — the conceptual representation.

their verbalization, their saccade rates increase. While this increase could have various causes (see Chapter 5.3.4), the epoch contrast is striking. Subjects first perceived and conceptualized the stimulus event (*message generation*) before fully preoccupying themselves with transforming it into language (*linguistic encoding*), and it was only then that they exhibited comparatively higher rates and lower travel distances. This plausibly supports the argument made in the previous passage on modality (cf. Ch. 5.2.2), namely that the representational modality is input-sensitive and may shift over time because of task demands. While LASS Theory and its precursors (Barsalou et al., 2008; Paivio, 1971) seem to account for such flexibility, they remain rigid regarding the temporal concurrence of modality-specific contributions to conceptual representation. While it seems plausible that a relative increase in saccade rate may indicate prominence of language-based as opposed to simulation-based conceptual representation, this would be a premature conclusion — how the frequency-measure of ocular activity might relate to situated simulation as opposed to language is further discussed in the context of the findings for H2 below.

### 5.2.4. Oculomotor suppression correlated with visualization intensity

Examinations of the concrete relationship between oculomotor activity and participants' self-reported visualization intensity were theoretically motivated, since the notion of sensory-motor enactment during cognitive processing is a disputed prediction of 4E cognition frameworks (Barsalou, 2008; Thomas, 1999). The eye-tracking evidence reported in Chapter 2.6.3 suggested the intuitive hypothesis that more reliance on situated simulation would express as more oculomotor activity, given that internal visual stimuli would be richer and trigger more and larger eye movements (Spivey & Geng, 2001; Spivey et al., 2000). As has been noted, none of these studies gathered data about their participants' experience of mental imagery and assumed that perceptual simulation would express as a homogenous phenomenon in their subjects (see Chapter 5.3.2 for further discussion).

The results of Hypothesis 2 contradicted this premise. Instead of a positive correlation, the present participants' self-reported visualization intensity was negatively correlated with oculomotor activity. Participants who reported vivid mental imagery exhibited marginally significantly lower travel distances and saccade rates, while participants with less vivid imagery displayed relative increases of rate and travel distance. Similar findings were reported in a blank screen study by Johansson and

colleagues (2011: 1203) and further discussed by Sima (2014). The remainder of this section discusses two different but related mechanisms that may have produced this inverse correlation. The first is derived from theories of attention and the second is taken from enactivist cognition.

First, why would strong visualizers reduce their oculomotor activity? Conceiving of the cognitive system as one that is information-processing and computational, situated simulations produce representations that can capture attention. While simulation does not project visible percepts onto the retina (although see Schwarzkopf, 2024), nor activate the peripheral sensors in a feed-forward fashion, simulations still require dedicated processing capacities in the visual cortex (Kosslyn et al., 2005). When visual simulations are drawn upon for conceptual representation, attention is automatically drawn inward to the top-down produced, transient simulation and no longer on the external world (Villena-González et al., 2016; Verschooren & Egner, 2023; Smallwood & Schooler, 2006; Brockhoff et al., 2023; Ehrlichman & Barrett, 1983: 20). Now, the participants in the present study kept their eyes open during the experiment, thus constantly feeding sensory impulses from peripheral sensors upward to the visual cortex, which naturally expects stimulation to continue[75] (see Chapter 2.5.2 for how vision and visual simulation run on the same neural substrates). Eye movements stimulate the retina and subsequently the visual cortex in a mechanistic fashion and do so regardless of whether external visual stimuli are informative or not, like a blank screen. According to Weiner and Ehrlichman (1976), any such eye movement interferes with internally produced visual representations, like the ones emerging from unconscious simulations[76]. Therefore, even spontaneous, non-visual eye movements impact the stability (or perceptibility) of a visual mental representation and consequently interfere with attentive processing of cognitively relevant information (Postle et al., 2006).

Such interference makes the present results seem all but unintuitive: Participants who succeeded at focusing interoceptive attention on the emerging simulations, particularly those who reported vivid mental imagery, automatically attenuated bottom-up stimulation by shortening saccade amplitude and lowering saccade rate (Kvamme et al., 2024: 3; for similar findings, see Demarais & Cohen,

[75] Even during fixations, tremor, drift and microsaccades keep visual stimulation above a minimum threshold to prevent Troxler fading (gradual saturation of cones and rods).
[76] See Chapter 5.3.4 for a discussion of relevant psychopathological evidence.

1998: 244). The attenuation of eye movement may have prevented disruption of emerging simulations in the visual buffer, thereby enabling them to become vivid and informative (see Chapter 5.2.4 for critical discussion). If this was the case, it would directly contradict tenets of enactive cognition, by which oculomotor activity is a causal requirement for mental visualization. If visual representation was the same as visual *pre*sentation to the mind (Fuchs, 2018: 133f.) — if interoceptive visual simulation was not qualitatively different from exteroceptive visual perception —, no difference between visualization intensities would be expected, as everyone is equipped with the same peripheral sensors to pick up their respective environment.

Second, did participants who moved their eyes a lot and gave lower visualization intensity ratings not activate situated simulations? Going by the above explanation, these participants would have had simulations still, but they may not have been able to focus their attention inward to a degree that would have allowed the simulations to become enriched visual representations (Kvamme et al., 2024: 3). None of the participants rated their visualization intensities 0 or even 1, asserting that they all became aware of some form of mental imagery. Considering the variability in mental imagery experience (see Schwarzkopf, 2024 for insightful discussion), the eye data of low-visualizing participants warrants a different explanation.

In enactivist theories, high-level cognition is understood as emerging from sensory-motor (i.e., low-level) interactions between the organism and its environment (Fuchs, 2018; O'Regan & Noë, 2001). Radical proponents hold, e.g., that verbs like *to kick* cannot be fully understood unless dedicated neural assemblies in the brain's motor system fire to produce corresponding, yet sub-threshold muscle activations in the legs. Enactivist conceptions posit that simulation in the visual cortex and concomitant phenomenology of visual mental images is caused by enacting eye movements that would typically occur during actual episodes of visual perception (Sima, 2014; Thomas, 1999; Spivey et al., 2000). Applying this logic to the present results of low-visualizing participants, launching eye movements aided their simulations of motion event space (Johansson et al., 2011: 1204) because their organism enacted typical efferent responses to visual stimuli, and the low-level behavior may have, in turn, excited their

visual cortex to a degree that it produced perceptible mental imagery of the events (Keogh, Bergmann & Pearson, 2020).[77]

The phenomenology of consciously experienced simulations in the visual system remains very much an open question (Schwarzkopf, 2024; Nanay, 2021; see Ch. 5.3.2.1). Asking participants to introspectively rate their experience on a unidimensional 5-Point-Likert scale (see Ch. 3.3.5.3.3) certainly did not do justice to the empirically confirmed variability (Blazhenkova & Kozhevnikov, 2009). From an analytical perspective, the different sample sizes of visualizers (see Table 4-7 in Chapter 4.4.7.2) equally constrain generalizability of the present findings. Consequently, it is a subject of further research to determine the conditions under which participants may either re-enact perception-based eye movements or attempt to block out external stimulation in their efforts to let simulations unfold (cf. Ch. 5.3.2 for further discussion).

The discussion about visualization intensity and eye movement patterns becomes more transparent when considering another central finding. The statistical models of H1 detected that movement direction effects, which constitute the crucial experimental condition to probe simulation in eye movements, persisted across all levels of visualization intensity[78]. This means that participants' eye movements were affected by the movement direction condition irrespective of whether they reported experiencing vivid mental images or not. On the one hand, this supports the argument that simulations were ongoing and impacted oculomotor characteristics tacitly and irrespective of individual variability. On the other hand, it supports the assumption that larger travel distances are a correlate of simulations of motion event space. If the magnitude variance of travel distances in the movement direction conditions depended more on visualization intensity, the movement direction coefficients would have returned insignificant in a statistical model that takes visualization intensity into account as a fixed effect. Since both predictors were significant, they both exert effects. While the levels of visualization intensity were negatively correlated with overall eye movement dispersion, the spatial extension of eye movements was still relatively

[77] Generally, enactivist accounts tend to be non-representationalist or non-mentalist, in that imagistic or otherwise quasi-sensory representations in the mind are superfluous and, even if they existed, were not used by cognitive processes. Mental activity can arise directly from bodily activity (Fuchs, 2018).
[78] This means that their coefficients returned significant in a model that had visualization as a predictor. So, the movement direction conditions still made a difference in a model where visualization intensity also explained variance.

enlarged by critical movement direction conditions, suggesting that the content of simulations modulates oculomotor behavior independently of attentional focus on simulations. In a metaphorical sense, this pattern is reminiscent of a 'gear-shifting' mechanism, whereby individuals with lower reported visualization intensity require a higher oculomotor 'gear' to effectively run mental simulations. As a result, when they exert cognitive effort (i.e., 'step on the gas'), their eye movements exhibit greater amplitudes, compensating for the low vividness of their mental imagery.

A similar argument could be construed from modeling saccade rate. While the full models did not detect a significant impact of visualization intensity on saccade rate ($p = 0.13$), a marginally significant decrease was detected for participants who reported strong visual imagery (Est. = -0.2; $p = 0.06$), and this became significant when analyzing only the encoding phase. A result thus limited in its significance has little empirical power and allows only for speculation, but it aligns with proposals of the previous discussion. In fact, repeated analyses of the Ehrlichman group (Weiner & Ehrlichman, 1976; Ehrlichman & Barrett, 1983; Micic et al., 2010; Ehrlichman & Micic, 2012) attest similar results: participants attending to visual imagery exhibited lower eye movement rates (as opposed to a verbal condition) to successfully gate out interfering visual input. Just as cognitive processing of internally generated images improves when eye movements are fewer (in relation to language-based processing, cf. Ehrlichman & Barrett, 1983), early neural responses to external stimuli are attenuated when attention is directed inward (Villena-González et al., 2016). Note that saccade rate was not affected by the movement direction conditions, suggesting a dissociation of the *process* of focused attention (or not) to simulations from their *content*, i.e., the simulated features of concepts. To conclude and return to the above question how the frequency-measure of ocular activity relates to situated simulation: How frequently participants move their eyes seems to be linked to their reliance on situated simulation as a medium for conceptual representation.[79]

Altogether, while it was not confirmed that travel distances enlargened proportionately with increasing vividness of visualization, this spatial measure of ocular activity is still affected by conceptualized space in motion event representations. Since predictable and systematic patterns show in travel distance data across all levels of visualization intensity, this underscores that simulation may be an underlying

[79] This paragraph only discussed why saccade rate decreased during attention to situated simulation. See section 5.3.4 for explanations why saccade rate would go up during verbalization.

computational mechanism in cognition that is applied to process different types of input and across individual differences. Saccade rate, though not straightforwardly associated, seems to be related to focused cognitive processing of self-generated internal representations, like emergent simulations.

### 5.2.5. Cognitive load affects saccade rates in encoding vs. recall

Related to the previous interpretation that saccadic rate could be a marker for attention to simulations in cognitive processing, another finding is worth examining. Oculomotor patterns shifted during recall phases, with participants showing more frequent but less spatially extensive eye movements than during encoding phases (— a »downscaling« (Johansson et al., 2018: 58); see also Brandt & Stark, 1997; Johansson et al., 2011; 2006; Barsalou, 1999: 591). These findings accord with predictions from theories of attention (Smallwood & Schooler, 2006) and event cognition (Radvansky & Zacks, 2014).

Principally, the two task phases in Experiment 1 required combinations of distinct cognitive processes to facilitate task completion. The encoding phase, i.e., on first encounter with the stimulus, demanded primarily perceptual processing, establishment of an event model, and subsequent encoding into memory (Zacks et al., 2007). During recall phases, perceptual processing was reduced to as much as was necessary for recognition of previously encoded, consolidated event models. Thus, establishing an event model is carried out more efficiently in the recall phase, due to lesser cognitive effort for automated processes of event perception, like segmentation and prediction error monitoring (Zacks et al., 2007; Radvansky & Zacks, 2014)[80]. Furthermore, the loop between bottom-up sensory input and top-down matching of event schemata required fewer cycles for the event model to settle, equally reducing cognitive load by minimizing cognitive change (Antrobus et al., 1964). The relative increases of saccade rate in the recall phases likely mark such shifts in balance from bottom-up sensory processing to top-down processing. Top-down event processing is driven by an established event model's predictions that return sub-threshold prediction errors — and thus do not require further, simulation-based elaborations, which are

[80] The corresponding event models are, in principle, already stored. However, the fundamental matching processes required for perceptual recognition still need to be performed again, albeit in an accelerated manner. Event perception in recall may be more selective because the cognitive system already has a set of candidate event models to choose from. Event models do not have to be reconstructed from scratch, but event recognition nonetheless, and at least, requires that a recalled event model quickly returns sub-threshold prediction errors.

assumed to decrease saccade rates (see above). In other words, since inward-directed attention to simulations became less important for event comprehension, eye movement control was released back into service of active external sampling, driving up saccade rate in the process (cf. Weiner & Ehrlichman, 1976; Villena-González et al., 2016; Smallwood & Schooler, 2006). Similarly, the comparatively lower saccade rates during encoding could be suggestive of simulation to be an important mechanism for how events are conceptually processed and encoded into memory (see H1). Conceptual elaboration of event models may be the cognitive system running simulations of event schemata (cf. pattern-completion inference, Barsalou, 2009), which may be necessary for a matching with the perceptual stream[81], and the lower rates might be indicative of increased reliance on simulation to support this schema elaboration and matching process.

Altogether, the relative increase of eye movement rate from encoding to recall could be interpreted in several ways. The comparatively lower processing demands in the recall phase allowed more efficient handling of cognitive resources for information processing, leading to quick task completion and return of the cognitive system to bottom-up sampling in preparation of the next trial, activating the external visual sensors and increasing saccadic rate (Antrobus, 1973; Verschooren & Egner, 2023: 1670). At the same time, the higher processing demands in the encoding phase required a more rigid gating-out of irrelevant bottom-up information sampling (from the blank screen) in favor of memory encoding, keeping saccadic rate low. An alternative account is compatible with findings by Micic and colleagues (2010), where cognitively driven eye movement rates are expected to be higher during information retrieval because it requires searching through memory networks, and lower when attention is directed at information currently held in working memory. It becomes apparent that there are alternative explanations for why saccade rate may change across experimental or subject-specific conditions. Exhibiting fewer eye movements might not be signaling attention to simulations exclusively but may also be related to cognitive effort (see Chapter 5.3.4).

### 5.2.6. Perceptual simulation supports message generation…

In Experiment 2, while movement direction effects did not hold for the entire trial period, horizontal travel distance was enlarged systematically in the epoch preceding discrete

[81] For instance, when one may have already recognized the event but not identified the acting entities.

speech planning. Like during the comprehension tasks analyzed in H1, these movement direction effects persisted despite variation in visualization intensities. In other words, when participants generated a message from perceived (non-verbal and) verbal input or to produce verbal output, their eye movements patterned in ways suggestive of situated simulations of motion event space.

One difference to the findings of Experiment 1 is that the eye movement patterns in the audio replay phase were not affected by the reported visualization intensities. While in H1, a negative correlation between visualization intensity and travel distance was found, no such effects were measured in Experiment 2. Regardless of visualization intensity, all participants were sensitive to the movement direction differences, which again supports a simulation account that is robust against individual variability regarding focused attention on visual imagery.[82]

The result that visualization intensities were not associated with eye movement patterns may be related to a change in task demands. Given that Experiment 2 demanded verbalization of the environmental sound events, participants may have adapted their apprehension of the perceptual input (i.e., already in the audio replay epoch) in ways more compliant with this demand. Possibly, they prioritized a verbal format of conceptual representation earlier on, distributing processing resources already during stimulus perception more evenly across *language* and *situated simulation* (LASS) for efficient language production. While they were handling sensory input in an encoding-specific fashion with representations constructed by situated simulation, in parallel, they were matching the resulting auditory objects with lexical concepts (Levelt, 1989), allocating processing resources to language to fast-forward conceptual processing for speech production — in psycholinguistic terms, participants already actively retrieved lexical units during stimulus perception.[83]

This contrasts with the comprehension task, where verbalization was not required and processing resources were available for actively attending to situated simulation, thus impacting oculomotor activity to larger degrees (H2). In the production task, although situated simulations seem to have occurred (cf. movement direction

[82] Demarais and Cohen (1998) similarly report an independence of directional effects on eye movements: »verbalization instructions resulted in an increase in horizontal eye movements relative to visualization instructions, without interfering with the effect of relational term on saccade direction« (p. 243).

[83] This also supports the view that conceptualization involves early activation of lexical concepts (cf. von Stutterheim, 1999), rather than being purely pre-verbal and uninfluenced by grammar or the lexicon (as posited by Bierwisch & Schreuder, 1992).

effects in audio replay), processing resources were also needed for linguistic encoding, and attention to simulations was likely attenuated, resulting in an absence of effects of individual visualization intensities on eye movements. In other words, whether one was a strong or weak visualizer did not matter as much in production as it did during comprehension, as simulations became backgrounded when conceptual representation is in service of linguistic encoding. Participants' eye movement patterns were not found to covary with their reported visualizations in Experiment 2, whereas they did in Experiment 1. A more symbol-based format of representation, like language, seems to be the preferred basis for conceptual processing during speaking.

### 5.2.7. … but speech planning remains unaffected

The interpretation that attention is captured to a lesser degree by situated simulations in the online cognitive processing for language-related tasks is further reinforced by the findings of Hypotheses 5 and 6. Movement direction effects were only found in the audio replay-epoch, i.e., before full attention turned to language production (H5), suggesting that the initial sensory processing tended to rely on situated simulations. On the other hand, even when construal, a crucial step in conceptualization for language production, resulted in an explicit verbalization of motion space, eye movement analyses showed no influence by movement direction during linguistic encoding (i.e., the pre-button/pre-voice epochs in H6). Taken together, when attention is drawn away from perceptual input analysis towards formulating an utterance, conceptual representation need not be enacted with corresponding eye movements, suggesting a backgrounding of situated simulations in favor of grammar-driven subprocesses of language production. This contradicts the recurrent thesis of 4E cognition frameworks that simulation constitutes a necessary mechanism for conceptual representation (Barsalou, 2008).

Assuming a less sensory-based conceptual representation is aligned with findings of H1, where the comparison between non-verbal and verbal comprehension showed systematic attenuation of oculomotor activity in the verbal modality. It neatly mirrors how eye movement patterns change when the representational modality used in conceptual processing is more perception-based or rather language-based. Saccade rates were significantly higher in the verbal comprehension task (as opposed to the non-verbal condition, Exp. 1, H3) and the epoch comparison in H5 confirmed similar increases in the pre-button/pre-voice epochs, where processing resources

converge on speech planning (as opposed to a reliance on situated simulation for perceptual processing in the previous audio replay epoch, Exp. 2). Therefore, verbal and non-verbal processing modes can principally be dissociated, and if they do operate in parallel, they may affect conceptual representation in a graded fashion (Barsalou et al., 2008).

In a similar vein, this difference also echoes assumptions discussed with respect to task demands of the memory experiment (H3), where increases in saccade rates were found for recall phases across both modalities. Proceeding from the audio replay epoch to the pre-button/pre-voice epochs parallels shifts in processing operations, moving from lower-level perceptual processing towards more effortful, higher-level cognitive operations (e.g., memory recall, subprocesses of linguistic encoding). More generally, participants' internal information sampling rate (cognitive change) rises when moving from one task-relevant operation to another (see Ehrlichman & Barrett, 1983: 23), independently of the specific demands of the memory task in comprehension (Exp. 1) or the verbalization task in production (Exp. 2).

Despite the activation of directional concepts during event construal, explicit verbal references to motion direction did not produce measurable changes in saccade rate or travel distance (H6), suggesting that overt reference to spatial elements does not directly shape oculomotor patterns during language production. This provides even stronger support for the assumption that speech planning activates more abstract, amodal representations and backgrounds the modal representations established by perceptual simulations (H5). All findings of Experiment 2 support the conclusion, posited i.a. by LASS Theory, that language processing may run primarily on linguistic, non-sensory representations — in other words, not every cognition is reducible to 4E principles.

### 5.2.8. Summary of the interpretation

Movement direction effects were present not only in Experiment 1 but partially in Experiment 2 as well, strengthening the above interim conclusion that the spatial measure of ocular activity was affected by conceptualized space in motion event representations. **Travel distance** was modified more strongly when conceptual representations were more sensory-based in comparison to those required for language-processing. Eye movement amplitudes seem to be influenced by the spatiality of the ***content*** that is processed (Spivey & Geng, 2001; Hartmann, Mast & Fischer, 2015: 5; Johansson et al., 2011).

This underscores an important claim about the simulation mechanism (see LASS Theory, Barsalou et al., 2008): since perceptual simulations affect eye movements even if they do not breach the surface of consciousness as vivid mental imagery, perceptual simulation is a routine operation in the conceptual processing of input — a novel finding is that situated simulation expresses more or less strongly in eye movement patterns as a factor of changes in task demands (e.g., cognitive effort) or of individual ability to maintain attention to simulations (visualization intensity).

People made fewer eye movements when their attention was drawn to internal visualizations, suggesting that simulation decreases **saccade rate** indirectly by shifting focus inward. Saccade rate also seems to be affected when the cognitive ***process*** itself changes, like in tasks that require parallel activation and integration of information from different levels of representation into a larger 'thought unit', such as during language production (planning an utterance, Exp. 2). In another scenario, the change in process may stem from completion of the current cognitive task and a subsequent return to active sampling (i.e., swinging open the bottom-up gate) in expectation of the next task, as occurred during recall trials in Experiment 1.

As posited by LASS Theory (Barsalou et al., 2008), *language* and *situated simulation* represent distinct modes of conceptual processing. The present findings show that they are dissociable during motion event processing in terms of behavioral expression: greater reliance on simulation corresponds with larger saccadic travel distances (across conditions) but lower saccade rates (across subjects), while greater reliance on language is associated with smaller travel distances but increased saccade rates.

These patterns support the view that simulation and language processing engage different attentional and oculomotor dynamics.

However, language and situated simulation do not share a monopoly on control of spontaneous eye movements. People who visualized strongly tended to exhibit both reduced saccade rates and shorter travel distances, suggesting that their eye movement patterns reflect an inward focus of attention on self-generated representations. This pattern is consistent with the idea that lower saccade rates mark attentional engagement with internal simulation processes. While the interpretation suggests an embodiment of information processing in non-visual eye movements, the temporal (rate) and spatial (travel distance) parameter of oculomotor activity may be differentially influenced by cognitive subprocesses: for instance, greater gaze dispersion might indicate the engagement of multiple modalities for a single task, whereas increased saccade frequency may signal heightened sampling or parallel processing demands (cf. Chs. 5.2.4 and 5.2.5). This speculative distinction would imply that time- and space-based oculomotor measures may serve as independent behavioral markers of distinct forms of cognitive engagement — a question that future eye-tracking research must tackle.

## 5.3. General discussion

### 5.3.1. Psycholinguistic limitations and implications

This subchapter discusses psycholinguistic limitations of the present analysis. First, the classification of verbal stimuli may not accurately reflect the experimental conditions they were meant to represent. Additionally, the incremental nature of language production complicates efforts to determine when spatial construal occurs during the verbalization task. Finally, participants may have relied on shallow *good-enough* representations that comply with immediate comprehension demands but obscure the deeper conceptualization processes targeted here.

#### 5.3.1.1. Interpreting verbalizations as motion events

Certainly, experimenter-bias stems from the classification of participants' verbalizations as motion or non-motion events (see Ch. 4.5.4 for detailed description of this procedure). Considering responses like *jemand schießt einen Ball gegen eine Wand* (someone is shooting a ball against a wall) to represent horizontal motion event rests on an important premise. The premise concerns Talmy's (2000b) proposal that motion events are necessarily conceptualized as representing movement along a path. While the above example is relatively informative because we can infer spatial information (source, path, and goal) from our knowledge about shooting balls and walls, other responses, like *jemand schießt Feuerwerk* (somebody is launching fireworks), do not foreground space as much. Classifying this response as describing a vertical motion event rests on the same premise, namely that the activation of a firework-shooting event schema necessarily activates spatial concepts that provide structure for simulated event representations.

Cross-linguistic examinations of motion event conceptualization (see Gerwien & von Stutterheim, 2022 for recent analysis), however, have debated the relevance of spatial components for motion event representation. While speakers of some languages indeed focus on representations of path or direction to categorize an event as one of motion, others are more focused on features of the moving figure and background the directional component of an objectively identical situation (von Stutterheim et al., 2012). Thus, the strategic decision to classify vague verbalizations as indicative of an underlying motion event representation where space was conceptualized is based on a premise that remains hypothetical.

The previous analysis of eye movement data suggests that, overall, the classifications were not as arbitrary as assumed. Stimuli classified as motion events triggered larger eye movements than stimuli characterized as non-motion events, licensing the interpretation that movement space affected non-visual gaze behavior, possibly through situated simulations of the encoded motion events.

#### 5.3.1.2. Incrementality limits analysis of epoch data

The interpretation of the results from the language production task is constrained by the serial-parallel nature of the involved processes (incrementality). The gaze data of Experiment 2 was labeled as belonging to distinct epochs, which are taken as rough temporal approximations to global steps in language production (message generation / linguistic encoding / articulation). Travel distances and rates were calculated epoch-wise. The principle of incrementality predicts that speakers may process different pieces of to-be-communicated information on different language production stages at the same time. For instance, while the figure-entity may have already been determined and prepared for articulation, lemma retrieval of a fitting verb to describe the event and finish the syntactic structure might be hindered due to disambiguation issues during preverbal message generation. While processing may have already yielded an articulable *topic* on which the rest of the utterance is going to *comment* (Krifka, 2008), other components have yet to reach lexical retrieval. Importantly, the participants in this study likely pressed the response key (to indicate readiness for articulation) as soon as the first element was articulable. Incrementality enables that some conceptual components may still be processed at a preverbal stage even after this button press, also even later during articulation of the first words. This strongly limits the interpretability of results for the epochs in Experiment 2 (H5, H6).

Not only is this epoch-classification a too simplistic operationalization of a highly complex process, incrementality restricts the interpretability of non-visual eye movements during language production experiments per se. This study assumed that event construal would be completed by the time participants pressed the button, marking a shift in online cognition to a more language-based form of processing. While this general shift in balance may in fact have been signaled by participants, this does not rule out that important processes of conceptualization (micro-planning: perspectivization, focus) had yet to finish (especially in the pre-button epoch). So, while the interpretation may sound convincing that situated simulations dominate

conceptualization during the audio replay epoch and become less dominant during the pre-button/pre-voice epoch, especially concerning the eye-tracking data (movement direction effects only in audio replay epoch!), it may well be that relevant conceptual components that would trigger movement direction effects on eye movements were only conceptually represented after the button-press.

Consequently, while participants' verbal responses reveal how they interpreted the environmental sound and although the epochal segmentation was by no means arbitrary (but precisely time-locked to participants responses, as close as you can get to their cognition without looking into their brain), incrementality makes it difficult to isolate a timespan in which participants dedicatedly conceptualized motion event space and, consequently, to argue that saccades were triggered by simulations of movement direction (see also Barsalou et al., 2008: 272 ff.).

The findings of H6, that even explicit verbalization of movement direction or goal did not correlate with larger travel distances in the pre-button/pre-articulation epochs, have an additional explanation now: Recall that in H6, motion event trials with explicit reference to space were compared to motion event trials without explicit reference to space. Incrementality allows for potential effect boosts of an explicit construal on eye movements to have occurred entirely in the audio replay epoch, as these directional components may have been conceptualized during early phases of preverbal message generation (perspective/focus), preventing such boosting from carrying over into later epochs.[84]

#### 5.3.1.3. 'Weaker' simulation in language-based conceptual processing

The present results question the assumption that the event representations underlying participants' comprehension of verbal stimuli are as enriched with perceptual simulations as their non-verbal counterparts. First, verbal stimuli induced comparatively smaller increases of travel distances, suggesting a defocusing of simulation representations in favor of assembly of concepts derived from lexico-syntactic structure. Second, it is likely that participants employed the strategy of attending to and encoding only the linguistic surface forms into memory despite

[84] In a similar vein, this limitation applies to the comprehension data, too. Spatial event components could have been conceptualized late in the trial or not at all, because participants were rather focused on entity-identification as opposed to merging identified components into an event-structure (see Ch. 5.3.3).

instructions to imagine the expressed event. Given the benefits of encoding specificity (Kent & Lamberts, 2008), some participants may have preferred to encode the verbatim linguistic stimulus and make the recall decision based on surface structures. While the sensory questions (cf. Ch. 3.3.5.2) were included precisely to prevent this, it is likely that participants merely encoded linguistic forms as opposed to constructing a simulation from a verbal event description. A central prediction of LASS Theory states that

> »when a word is perceived, the linguistic system becomes engaged immediately to categorize the linguistic form (which could be auditory, visual, tactile, etc.). [… W]e assume that the linguistic system and the simulation system both become active initially, but that activation for the word form peaks before activation for the simulation« (Barsalou et al., 2008: 247).

In other words, simulation-based conceptual processing of language is not a question of 'if' but of 'when'. At the same time, however, Barsalou and colleagues (2008: 268ff.) acknowledge that under specific circumstances (see Truman & Kutas, 2024 for extensive discussion), such as when conceptual elaboration is not required, language-based conceptual representation is sufficient to handle cognitive tasks. This is not a new idea, and so-called *good-enough representations* (Ferreira, Bailey & Ferraro, 2002; see Frances, 2024 for review) convincingly explain how even in information-dense conversation, interlocutors achieve mutual understanding despite incomplete, temporarily incorrect, masked, or missing bottom-up information. The good-enough approach to sentence processing provides a plausible explanation for why participants in this study processed verbal stimuli for a shallow understanding and did not delve into situated simulations for conceptual representation: As soon as participants arrived at a good-enough understanding of the sentence, they stopped conceptual elaboration and encoded the verbal form, or even just isolated lexemes.[85] How utterance meaning may be represented conceptually in a language-like format (Kintsch, 1988) was discussed at length in Chapter 2.4.1. Speed and colleagues (2015: 199) sum up that »[s]imulation may not be necessary for shallow language tasks, where a good-enough representation could be inferred simply from linguistic information alone, using statistical relations between words.«

[85] A related trend that deserves mention is that some participants produced predominantly one-word responses in the verbalization task. Isolated, nominalized responses do not constitute events per se – because they lack the temporal dimension – but participants had to have processed a bare minimum of the environmental sound event to retrieve a fitting verbal label.

Likewise, Truman and Kutas (2024: 12) assert that reliance on perceptual simulations is context-driven »and may not be necessary for word comprehension in tasks that do not require attending to those features.« In analyses of EMG data, for instance, Foroni (2015) found that facial muscle impulses of advanced foreign language learners were weaker in a foreign language vs. their native language when they processed 1st person predicate verb phrases with lexemes like *smile* or *frown*. Effectively, subjects processed utterance meaning in the foreign language in a less embodied fashion. The absence of comprehension issues, however, attests that the subjects constructed a conceptual representation based on language and without comparable contributions of situated simulations. In other words, the linguistic system sufficed for meaning construction (Kintsch, 1988). Comprehending a foreign language may, in general, rely more on a verbal processing mode, underscoring the value of foreign language acquisition research for examinations of concept embodiment (Pavlenko, 2012; Abbassi et al., 2015). Contradicting predictions from LASS Theory, situated simulations or deeper conceptual processing may not always be required when conceptual representation is language-based.

Note, however, that the findings of Foroni (2015) — where muscle responses were simply weaker and not absent — along with the present results do support the involvement of simulation in conceptual processing, even if its contribution varies by task. If simulation had been entirely absent during the verbal memory task (Exp 1, Part 2), for instance, movement direction effects should not have occurred at all. However, such effects remained significant, although less pronounced than during the environmental sound memory task, indicating that simulation processes were attenuated but not completely deactivated. This may be, as mentioned above, due to participants flexibly turning to situated simulations for conceptual representation due to task demands and, ‘when’ they do, in a weaker form. The different modalities for conceptual representation may be attended to or activated in a graded fashion, and not in an all-or-nothing manner — contradicting strong claims that sensorimotor simulation is a no-matter-what process in cognition (Barsalou, 1999; Fuchs, 2018).

### 5.3.2. Controlling the content of perceptual simulation

Participants could have simulated anything that spontaneously came to mind during the task. Most certainly, not all participants consistently activated the same motion event components, especially those that Talmy (2000b) hypothesized to be essential, such as path direction (cf. Ch. 2.6.2).

Departing from the rejection of H6, one might wonder whether participants' verbalizations really revealed the content of their conceptual representation, whether their motion event simulations were structured by idealized path representations (Talmy, 2000b), and whether it was the activation of path representations that triggered the eye movements. Accordingly, it merits discussion whether participants' perceptual simulations consistently evoked the visuospatial path representations that were presumed to be necessary for constructing the meaning of motion events.

This section is structured as follows. First, a short review of empirical evidence confirms extreme variability in mental imagery, its nature, content, and experience. Second, language is presented as a means of gaining some control over the content of simulations. Third, the ventral-dorsal processing streams are discussed as the cortical structure that underlies eye movements based on motion event simulations.

#### 5.3.2.1. Variability in visual mental imagery

Visual mental imagery is characterized by extreme variability (for review, see Nanay, 2018; 2021).[86] This variability applies not only to the vividness or level of detail in the experience, but also to its content and structure (Barsalou, 2009: 1282). Unlike perceptual experiences, mental imagery is not constrained by the same structural rules and can be manipulated in ways that visual percepts cannot. For example, mental images can be resized, reshaped, or reoriented. We can shift our mental focus, scan across imagined scenes, or stack and overlay elements (Kosslyn et al., 2005, Barsalou, 1999: 591; Laeng et al., 2014). Simulations may be static or dynamic, exhibiting phenomena such as representational momentum (Freyd & Finke, 1984; Huber & Krist, 2004). They may occur voluntarily or spontaneously (Nanay, 2021; Smallwood, 2013). Mental images are neither universal, i.e., not everyone experiences the same image for the same concept, nor consistent within a subject, i.e., the same

[86] This section mainly concerns mental imagery in the visual modality. See Nanay (2018), Floridou et al. (2021), or Lacey and Lawson (2013) for an overview of mental imagery in other modalities.

concept may conjure various images across different instances in the same person, because simulations are highly context-dependent and flexibly adapted to situational demands (Truman & Kutas, 2024). There are qualitative differences in the richness, clarity, and detail of mental images, as well as how people 'perceive' them (Schwarzkopf, 2024). However, there tends to be cross-modal consistency within a person, e.g., individuals with strong visual imagery exhibit strong auditory imagery (Floridou et al., 2021). And despite individual variation regarding excitability (Keogh et al., 2020), the human visual cortex generally shows heightened activation whenever visual representations capture attention (Somers et al., 1999).

One dimension of the variability of mental imagery concerns the perspective that subjects take on situated simulations of events. Perspective is understood as the vantage point of the mind's eye[87] on simulated visual space and therein arranged entities. While it is generally accepted that there is not a default or universal perspective from which events are viewed (Chatterjee et al., 1999: 401; Tversky, 2005), subjects may adopt an egocentric, first-person (e.g., Glenberg & Robertson, 2000; Stanfield & Zwaan, 2001), or an allocentric, third-person perspective on a mentally unfolding simulation (Spivey & Geng, 2001). Perspective appears to depend on recency and frequency of exposure (Zwaan & Madden, 2005), suggesting that prior experience and contextual salience are influential. Radvansky and Zacks (2014: 70f.) propose that language may offer a means of stabilizing variability in perspective, providing a potential means of gaining some experimental control over the content of simulations.

#### 5.3.2.2. Language

Verbal stimuli have been used frequently in studies of embodied language comprehension (see Zwaan & Madden, 2005). While many have focused on simulations triggered by individual words or word pairs, others have used short, decontextualized sentences that describe events (see Chapter 2.6.4). One simple way of influencing perspective on events is to modify grammatical categories like person or mood (Talmy, 2000a; 2000b; Klein, 2009). First-person pronouns in verbal stimuli induce egocentric perspectives in simulation, and third-person pronouns tend to trigger allocentric representation (Glenberg & Kaschak, 2002; Vukovic & Williams, 2015). This

[87] Pylyshyn's (1973) seminal critique of Kosslyn's imagery theory was titled *What the mind's eye tells the mind's brain*. Popular US-American neuroscientist Oliver Sacks (2010) published a best-seller entitled *The Mind's Eye*. Johansson (2013) named his PhD thesis *Tracking the Mind's Eye*.

is important in the present context, since the eye movement patterns here were assumed to be triggered by taking an observer-perspective. Accordingly, all verbal stimuli were designed as sentences with a third-person subject noun phrase.

Another grammatical category that has aroused research interest in this context is aspect, which modulates the internal temporal structure in described events. In English, for instance, imperfective or progressive aspect is used to characterize a situation as currently ongoing or unfinished relative to the time of speaking (or topic time; Klein, 2009). Perfective or simple aspect construes events as completed. Accordingly, processing of imperfective aspect has been shown to steer focus in simulations on ongoing, dynamic processes in events (Bergen & Wheeler, 2010). Huette and colleagues (2012) studied spontaneous blank screen eye movements and found that gaze dispersion, which is conceptually similar to travel distance, was significantly larger when participants comprehended event descriptions encoded with progressive vs. simple aspect. Considering that German lacks a grammaticalized aspectual distinction of this sort, however, it remains uncertain whether the participants in this study conceptualized the verbal motion event stimuli as ongoing or holistic, i.e., completed.

Critically, von Stutterheim and colleagues (2012) reported a language-specific preference of their German sample to visually attend to video stimuli in a rather holistic fashion. While exteroceptive visual processing is less of concern here, their findings point to an important effect of language on event conceptualization, hence simulation. Subsequent research (for review, see von Stutterheim & Gerwien, 2023) consistently supported the assumption that language-specific lexicalization patterns of motion events differentially profile conceptual components in motion event schemata. For example, while verbs like *s'avancer* (French for to advance oneself, reflexive) foreground orientation and deictic direction of the moving figure, German *rennen* (to run) emphasizes the manner of figural motion and defocuses direction. Language-specific motion event schemata are characterized by »assignment of weights to specific conceptual components« (Gerwien & von Stutterheim, 2018: 235). As such, when German speakers activate motion event schemata already during preverbal message generation, i.e., while establishing event models, specific conceptual components of motion may be more salient than others, which may induce variability in how they simulate the situations they picked up from the sensory stream.

When words activate concepts, they trigger sets of abstract categorical features that constrain simulations to the most characteristic attributes (Lupyan, 2012; Lupyan & Bergen, 2015: 415), even perceptual features like shape or color. Moreover, this labeling function extends beyond individual words to the broader hierarchical structure of semantic memory. Superordinate categories pass down generalized features (e.g., direction of motion or body orientation in locomotion), which remain available unless overridden by more specific sensory input (Anderson, 2013: 106-107). In fact, during online conceptual processing, the surface forms of the input can defocus these inherited features in favor of more salient ones, such as the manner rather than the path of motion, which could additionally explain why the directional axis — since it is associated with a superordinate schema — may not have been activated as much by the verbal stimuli as by environmental sounds — where it is a feature of the input. Comprehenders have options to conceptualize objective situations in different ways, and language may bias them to prefer one over another.[88]

Notably, it is not only lexical items that serve this labeling function. Talmy (2000a: 29) emphasizes that grammatical morphemes also contribute to how events are represented, though in a slightly different way: rather than referencing concrete objects, they specify abstract relational structure between entities (cf. Landau, 2017; Langacker, 1986).

Given that verbal labels activate both abstract and modality-specific representations of concepts (Lupyan & Thompson-Schill, 2012), labeling is a central way in which language modulates the content of simulations. At the same time, it is also what sets it apart from how simulations arise from environmental sound processing. Whereas language restricts the simulation to category-specific attributes, environmental sounds represent events in an already concrete and specific fashion, affecting simulations intrinsically through the audible perceptual details (cf. Chapter 2.3). For instance, simulations activated by a verbal stimulus like *jemand joggt* (someone is jogging) may be enriched with inferences about how jogging typically takes place outside. An environmental sound of someone jogging on a gravel road, though capturing the same abstract idea, causes potential simulations to be more specific regarding this schematic component (ground). Similarly, controlling for perspective is not as straightforward with environmental sounds. Since environmental

[88] The next section on ventral vs. dorsal input processing suggests how this conceptual flexibility may be neurally implemented.

sounds appear as quasi-sensory, vicarious experience, it is up to the participant which perspective they assume[89], unlike in verbal stimuli, where perspective can be specified through grammar. Concerning grammatical aspect, however, it is out of the question that environmental sound motion events are perceived as ongoing[90], whereas this is left implicit in German utterances.

Altogether, language prescribes abstract, category-specific features to simulations, leaving it up to subjects to enrich this representation with additional perceptual details. Environmental sounds are by default perceptually predefined instantiations of concepts, i.e., experiential tokens of a category, and thus naturally lead to more contextually grounded and detailed simulations.

### 5.3.3. Dorsal and ventral pathways in motion event simulation

The documented variability in the phenomenology of mental imagery casts doubt on the hypothesis that conceptualizations of motion events necessarily activate simulations of path-direction information (see Chapter 2.6.2), and the aforementioned labeling (Lupyan, 2012) or *profiling* (Langacker, 1986: 6) narrow the myriad simulations that could be activated to represent any given event conceptually. Neurophysiological findings suggest that the degree to which concepts, irrespective of how they are activated, become associated with sensory or behavioral responses (like eye movements) depends on frequency and context of experience with the concept (for review, see Hauk, 2016; Truman & Kutas, 2024). With respect to conceptual representations expressing in oculomotor behavior, Liu (2009) argued that attention was the driving factor because it mediates the activation strength of the neural substrates in sensorimotor systems that are responsible for oculomotor control. While Liu's (2009) conclusion concerns the psychological foundation of motion-based eye movements, neurological findings suggest that oculomotor dynamics change when cognitive processing requires differential activity on the neural level, namely in the dorsal or ventral pathways. This subsection discusses how these dedicated neural

[89] Note however, that the placement of the microphone during recording of an environmental sound may influence, which perspectives may be taken on the audible situations.

[90] In horizontal environmental sound events, the movement path itself is directly audible, whereas in vertical events (e.g., *fallen*) the path is typically inaudible. This difference likely makes the movement trajectory more salient in horizontal events, thereby exerting a stronger influence on eye movements. By contrast, vertical events primarily convey the goal of motion (e.g., impact on the flow), so any representation of the path must be inferred, whereas in horizontal events the path is perceived rather than inferred.

pathways may distinctly shape simulation-based eye movements and how deeper understanding of these structures may help interpret non-visual gaze data.

The dorsal and ventral pathways are an integral part of a cortical structure that hosts important functions for visual processing. In general, when an external stimulus is perceived, it is processed in the ventral stream for identification and recognition (i.e., for 'meaning') and simultaneously be localized in visual space through the dorsal stream. The ventral stream (also called the *what*-system) supports identification of visual percepts and the dorsal stream (the *how*-system) helps us localize these visual percepts and prepares us for moving towards and acting upon them (Mishkin, Ungerleider & Macko, 1983; Milner & Goodale, 1995).[91] Importantly, language comprehension relies heavily on activity in the ventral pathway (Hickok & Poeppel, 2007). At the same time, the dorsal stream is involved in the classification of shapes and sizes (Konen & Kastner, 2008; Zachariou et al., 2014), which are features that support identification and recognition, suggesting that the involvement of both streams in visual processing is not mutually exclusive but converges in specific tasks (see also Deubel et al., 1998; Landau, 2017). Logie (2003) and Zimmer (2008) conclude from comprehensive reviews of neurocognitive research that, while resources are distributed to both at the same time, we trade off separate working memory capacities to represent space vs. object identity due to anatomical segregation of ventral and dorsal pathways.

Accounting for this distinction, the *Object-Spatial Imagery and Verbal Questionnaire* (Blazhenkova & Kozhevnikov, 2009) contains sections that specifically assess imagery abilities concerning spatial relationships or visual features of objects. Johansson and colleagues (2011: 1203) found that blank screen eye movements during recall of a visual scene were shrunk in participants who observed high spatial-imagery scores, while higher ratings of reported vividness of visual images correlated positively with scores on the object-imagery questions. This is not surprising, as eye movement control is located in the dorsal pathway (Colby & Goldberg, 1999), and the dorsal prefrontal cortex (Zimmer, 2008: 1383), subserving functions like cognitive and action control, is assumed to cache spatial coordinates for targeted execution of eye movements.

[91] Auditory processing also relies on the distinct contributions of ventral and dorsal processing (Bizley & Cohen, 2013; Zimmer, 2008).

The remainder of this section builds on this distinction and speculates how activation strength in the different pathways may be linked to variation in oculomotor activity. Stronger activity in the dorsal stream during motion event processing is assumed to lead to more pronounced eye movements. In reverse, more pronounced eye movements are taken as a signal for increased activity in the dorsal stream, indicating that internally generated visual representations were based on space rather than on object features. In contrast, when there is stronger activity in the ventral stream, the internal visual representation was likely based on object-features that helped identification or recognition, and fewer, less pronounced eye movements should occur.

This distinction is in line with findings on topology in mental representation that suggest that discrete localizations of objects (e.g., prepositions like ‘in’, relating to the ground-object properties) are differently represented from axes, because axes imply relative locations (prepositions like ‘above’, where a figure is related to a ground) (Carlson-Radvansky & Jiang, 1998; see also Estes et al., 2008; Spivey & Geng, 2001; Demarais & Cohen, 1998; cf. Ch. 2.6.3.2) and induce processing reliant on the dorsal stream, which hints again at a biasing of perceptually simulated content towards space: »when we talk about places, we tend to ignore attributes, parts, or functions of the object being located« (Carlson-Radvansky et al., 1999: 516). Sima (2014), interpreting findings of Johansson and colleagues (2011), proposes accordingly that »people with a high spatial mental imagery score will mentally imagine much less visual information, e.g., shapes, textures, than a person with a low spatial imagery score« (Sima, 2014: 107), reiterating that one either leans more towards object-based or space-based imagery.

The findings converge on the idea that neural processing and mental representation of objects in visual space (dorsal pathway) is different from that of object-specific visual features (ventral pathway) and that either mode differentially affects eye movement patterns. Concerning the experiential domain of motion, Wu and colleagues (2008) report pertinent neuroimaging results that align with this distinction of object-identity vs. spatial relationship. When participants’ attention was captured by path information in video clips of motion events, activity was heightened in the dorsal pathway, suggesting processing of the relative location of the moving figure in the video. When participants were attending to manner information, increased activation was measured in the ventral stream, indicating focused processing of characteristics inherent to the figure-object. Therefore, manner of motion is a property of the figure,

requiring closer attention to a specific object, and path of motion is grounded in space, recruiting neural processing associated with relative spatial shifts of mentally represented entities. Contradicting proposals by Talmy (2000b; see Ch. 2.6.2), people do not always think about figures moving through space when they think about motion events, because they may also conceive the figure with particular features of its moving body without necessarily considering that the figure's movement is physically bound to a relative space.

Now the central question of this subsection can be answered: Does activation of path representations during motion event simulation trigger eye movements? Systematic, path-reenacting eye movements during motion event processing should occur if the motion event was conceptualized as constituting a spatial relationship of a figure-entity to a ground-entity (i.e., representing the figure as a participant in an event of movement, which per se requires space in which movement unfolds and in reference to which the figure-object is arranged). Path-reenacting eye movements are less likely to occur when motion conceptualization concerns a modification of the figure-object representation (i.e., as if motion was something that characterized the figure, revealing that it moves in a specific manner, therefore defocusing the translatory aspect) (see Spivey & Geng, 2001: 240). When eye movements systematically show directional effects derived from motion events, it is likely that the dorsal pathway was activated in response to a mental representation of the motion event which, at least *to some degree*, was specific about the spatial relationships between the figure and the reference ground on which the figure moved (see Ch. 2.6.1). This opens one plausible window to observe conceptual representation of motion events through systematic eye movements: when eye movements enact path space, dorsal neural processing of path was likely integral to thinking, suggesting that path information was salient in conceptual representation.

Why only 'to some degree'? As mentioned above, dominant activity in the ventral stream (object-identity ~ figure-based information) does not block concurrent processing in the dorsal stream (space ~ path information) — rather it is gradient (Zachariou et al., 2014). The results of H1, H2 and H4 support such a gradient. Both in the comparison of non-verbal vs. verbal processing, throughout all levels of visualization intensities, and before participants fully engaged in linguistic encoding, the motion event stimuli correlated with increases of travel distance due to movement

direction conditions (H1). High-visualizing participants, while possibly focused on representing object-identity, exhibited fewer eye movements (ventral processing), but the few eye movements they made were still affected by motion space (suggesting concurrent processing in the dorsal pathway). The movement direction effects held across all levels of visualization intensity self-ratings (H2) and low visualizers likely had a proclivity to process motion event space (dorsal stream)[92] as opposed to object detail, thus exhibiting larger oculomotor responses while still retaining the movement direction differences. Plus, even when the verbalization task required rapid identification (ventral) of auditory objects for activating lexical concepts, movement direction effects were present, suggesting again parallel activity in the dorsal pathway (H4). The insignificant results of H6, however, become more puzzling: even utterances that contained overt references to both movement direction and goal did not boost eye movement, although conceptual representation of spatial relationships is supposed to trigger dorsal stream activity.

#### 5.3.3.1. Task-induced ventral vs. dorsal processing

Did task instructions induce that participants mentally represent object (ventral) versus spatial (dorsal) information? First, the memory task inherently biases participants toward object identification and elaboration, as they were specifically instructed to recognize and remember the events. This likely caused focus on 'what'-information, drawing attention to audible object details and a preferential encoding of central protagonists rather than integrating them into an event structure ('how'). Second, the verbalization task promotes identification processes, since participants must retrieve verbalizable concepts quickly to initiate speech planning.

The interpretation of the results may be relativized if task demands created preferences for cognitive processing associated with ventral stream activity. For instance, introspective assessments of visualization intensity ("How vivid…") may have primarily elicited judgments about visual object imagery or object details and only indirectly revealed participants' spatial imagery skills. Johansson and colleagues (2011: 1203) found that when »participants are to rate their own visualization, this judgment is primarily associated with vividness and object imagery aspects«. On top of that, the analyses of Johansson and colleagues (2011) and Kozhevnikov and

[92] See the following subsection (Ch. 5.3.3.1) for explanation (Johansson et al., 2011; Kozhevnikov et al., 2010).

colleagues (2010) suggest a trade-off, that is, individuals who scored well in spatial imagery tasks did not perform well in visual object imagery tasks and vice versa. Therefore, data from high visualizers might need to be reframed as indicative of object-based (ventral) rather than space-based (dorsal) imagery, which is, speculatively, what the low-visualizers of this study should be good at.

However, as the results suggest, both processing streams were engaged concurrently, with a shifted emphasis toward ventral processing (cf. Zachariou et al., 2014). This shift may have affected the spatial magnitude of direction-dependent eye movements, rather than their occurrence per se. Additionally, the supposed predominant activation of the ventral pathway by the verbal memory task may inform the interpretation of the travel distance data: since verbal stimuli contain concept labels, preference for ventral processing may have developed, and a corresponding backgrounding of dorsal engagement may have led to weaker activation of the oculomotor system (resulting in 'staring') and reflect a stable mental simulation of object features over spatial dynamics.

#### 5.3.3.2. Inferring processing mode from behavioral measures

Independently of eye movement data, how else could it be assessed whether participants relied preferably on situated simulations or on verbal processing during event conceptualization? One possibility is to compare recall performance. Participants who relied on verbal labels might be faster and more accurate during stimulus recognition, exhibiting lower error rates overall, and additionally committing fewer errors in the verbal memory task compared to the environmental sound memory task. Furthermore, if participants engaged in inner speech that concerned the stimulus during the non-verbal memory task, they are likely to start speaking faster after the button press and to articulate more fluently during the verbalization task, making fewer hesitations and repairs, since they pre-activated lexemes through using inner speech. These assumptions remain speculative and would require targeted future research to be empirically tested.

#### 5.3.3.3. Summary

Altogether, the processing of motion through space- versus motion-related characteristics of moving figure may be differently weighed towards different cortical pathways, each of which distinctly recruits the oculomotor system. Predicates that

make manner of motion salient primarily trigger a figure-focused conceptual representation as opposed to propositions about an entity's movement trajectory relative to a ground, causing differences in how strongly motion event simulations foreground spatial conceptual components and lead to expression in overt eye movements. Future studies that triangulate neuroimaging and eye movement data recorded from carefully designed motion event stimuli will be necessary to illuminate the speculative relationships described in this section.

### 5.3.4. Eye movements as markers of cognition

In the discussion so far, a few arguments are based on links between non-visual oculomotor activity and cognition. Most notably, a distinction was made between the meaning of a spatial measure of oculomotor activity (saccadic amplitude, here operationalized as travel distance by trial) and that of a frequency measure (saccade rate). This section discusses how spontaneous oculomotor activity can reliably and informatively tell us how we generate representations from knowledge.

In this study, travel distance was interpreted as an analogue expression of spatiality in mental representation. It is viewed as being directly influenced by characteristics of a mental simulation. However, the absolute amplitude of individual saccades varies strongly not only within but across individuals and also seems to be dependent on how readily someone's attention was captured by internal visualizations of event representations. Saccade rate, on the other hand, appears to have been affected by how strongly someone's attention is drawn to mental images emerging from situated simulations. Saccade rate was therefore interpreted as a marker of focus on situated simulation for conceptual processing. Relative to the participants' observed frequency of eye movement, lower rates would mean that simulation generated visual representations that were 'attendable' or perceptible enough to decouple the eyes (the afferent, peripheral sensor of the visual system) from exteroceptive sampling. Relatively higher rates would indicate the opposite, that the emerging simulations were not powerful enough to induce an inhibition of oculomotor activity.

How, in the present study, do both the spatial and temporal eye movement measures link up to mark that cognitive computation is busy generating simulations? The stronger the object-based simulation, the lower the saccade rate and the smaller the relative travel distance of the strongly visualizing participants (see Johansson et al., 2011; Kozhevnikov et al., 2010). In contrast, the higher rates of the weakly visualizing participants indicate less reliance on object-based simulation for conceptual representation. Yet, even when the object-based representations were less vivid, the spatial influence on travel distance is retained for low-visualizing participants.

At the same time, Experiment 2 (H5) casts doubt on the informativity of eye movements for characterizing simulation. Results from the verbalization task demonstrate an increase of rate *without* condition-specific increases in travel distance. This suggests

that saccade rate could also be a marker of how language-based someone's processing was, contradicting the above interpretation: saccade rate would instead mark attention capture more verbal versus simulation-based conceptualization (cf. LASS Theory; Barsalou et al., 2008). But why would saccade rate go up when conceptual processing becomes more verbal?

First, if focus on visual representations need not be sustained for the production of words, decoupling the eyes from exogenous perception would not be expected to improve the speech planning process. Keeping the eyes still and letting simulations generate conceptual information undisturbed may automatically further elaborate the to-be-encoded message with information that may neither be required nor useful for efficient utterance production (Weiner & Ehrlichman, 1976: 41). Keeping the eyes still would invite unsolicited generation of conceptual components on the message level, which may interfere with fluent linguistic encoding because these unnecessary conceptual components are routed into already ongoing linguistic encoding. This would interfere with completion of a cohesive syntactic plan and result in hesitations or repairs during articulation. This interpretation suggests that keeping oculomotor activity low, i.e., maintaining low saccade rates, provides a proprioceptive bottom-up cue for the conceptual system to engage in mind-wandering (Smallwood & Schooler, 2006; Foulsham et al., 2013). In this sense, eye movement interferes with the generation of perceptual predictions (Radvansky & Zacks, 2014; Hobson, Hong & Friston, 2014) and the concomitant processing of thus emerging visuospatial mental images (Postle et al., 2006; Shapiro, 1989; Wilson et al., 2018). Moving the eyes during linguistic encoding may be a behavioral routine to accompany efficient language production, aiding the utterance planning by preventing unwanted elaborations on the level of conceptual representation.[93]

Already described above, a compatible argument concerns incrementality. The serial-parallel mode of operation of the language production process is a multi-faceted, high-load task that entails frequent cognitive change (Antrobus et al., 1964: 245; Ehrlichman & Barrett, 1983; Micic et al., 2010; Marconi et al., 2023) between operations necessary for language production: online information selection and structuring, lexical retrieval, grammatical, morphological, and phonological encoding, articulation and output monitoring. Moving the eyes might help individuals handle this

[93] This is compatible with modern accounts of the brain as a predictive, hypothesis-testing system that continuously updates its internal model of the world to minimize prediction error (Friston, 2012).

highly automatized but complex speech production process. So, while keeping the eyes still invites the conceptual system to continue elaborating a message, moving the eyes binds resources to make the effortful integrative process of utterance production more efficient (Weiner & Ehrlichman, 1976).

Third, and following the above discussion about activity in ventral vs. dorsal pathways during conceptual processing, it is tempting to assume that language-based conceptualization (cf. LASS Theory) would co-occur most strongly with activity in the ventral stream. Recall that the ventral stream is engaged in identifying and recognizing both perceptual and linguistic information, such as referents in utterances (Hickok & Poeppel, 2007). If increased activity in the ventral stream 'trades off' resources away from dorsal stream processing — where heightened activation co-occurs with increased oculomotor activity —, why then would language-based conceptualization be at all related to increases in eye movement patterns? Language can still evoke spatial simulations or attentional shifts that still lead to increases in eye movement patterns. It is likely that Barsalou and colleagues (2008) are correct in their assumption that, although language-based processing peaks first, both language and situated simulation become active at the same time, they just peak in activity in a delayed fashion, potentially due to the availability of cortical processing resources.

#### 5.3.4.1. Attempting to generalize ocular markers

At this point, it becomes clear that interpreting non-visual eye movements as markers of cognitive activity is fundamentally constrained by the specific demands of the task. This is nothing new, and it is generally accepted that there is not a one-to-one relationship between the multi-dimensional thought that humans are capable of and the controllable parameters of ocular activity (blink, movement in space, movement across time). Proposed dictionaries for the interpretation of eye movements in cognitive terms (originally by Bandler & Grinder, 1979; see Diamantopoulus et al., 2009) have not withstood empirical falsification (Marconi et al., 2023). The following discussion concerns whether the interpretations of the eye movement data of the specific tasks in this study can be corroborated with or extended to findings from non-visual gaze research in other disciplines.

As has been mentioned, one view interprets the reduction of eye movements in non-visual gaze as indicating activity of situated simulations in conceptual representation, which is supported by empirical research on mind-wandering (Antrobus

et al., 1964; Smilek et al., 2010; Villena-González et al., 2016; see Steindorf & Rummel, 2020). Another view associates non-visual gaze with cognition in a mechanistic fashion, linking the occurrence of spontaneous eye movements less to contextual specifics like motion event space, but to domain-independent cognitive activity, like long-term memory retrieval (Ehrlichman & Micic, 2012), cognitive load (Vredeveldt et al., 2011), decision-making (Gold & Shadlen, 2000), or verbal vs. image-based processing (Ehrlichman & Barrett, 1983; Liu, 2009).

A chance finding of the analysis may contribute to this empirical issue. A result of Hypothesis 1 (cf. Ch. 4.4.7.3) confirmed that saccade rate was significantly associated with experiment duration. Rate of spontaneous eye movements increased linearly with time spent on the experiment (Est. = 0.002[94], $p < 0.001$). Thus, the same participant was more likely to make more saccades in the last trial than in the first trial. This might be because, over time, participants became more fatigued and progressively lost concentration. Therefore, a more general observation emerges: Frequency of non-visual saccades may be negatively correlated with participants' ability to concentrate. In a way, this may explain not only the effects of trial sequence on rate but could align neatly with the claim that more vividly experienced visual imagery reduced saccade rate. Participants' attention must be focused on the internally generated image to retain its vividness and extract task-relevant information, and saccade frequency would signal how successful they were in doing so. While this may be a plausible explanation[95], the present study did not analyze this hypothetical connection and must point to further research on non-visual gaze for resolution.

Past research in psychology, however, strongly supports that reduced ability to focus on a unique, self-generated thought is associated with changes in eye movement patterns. This finding has aided psychopathological diagnoses as well as the development of psychotherapeutic methods. For instance, patients diagnosed with attention-deficit/hyperactivity disorder (ADHD) have difficulties maintaining fixation on a percept and have overall shorter fixation durations (Siqueiros Sanchez et al., 2020; Fried et al., 2014; Munoz et al., 2003) — a behavior which characteristically aligns with

[94] Over the course of 96 trials, saccade rate increased by 0.19 on average. Since the coefficients were calculated on the basis of a transformed variable, it is difficult to interpret this number. All one could say is that relative to the first trial, in the last trial saccade rate was increased to a significant degree.

[95] Admittedly, it is far-fetched to assume that low-visualizing participants, who moved their eyes more than high-visualizing subjects, were not able to concentrate during Experiment 1, let alone that engaging in linguistic encoding (Exp. 2), which also drove up eye movement rates, would somehow be connected to diminishing concentration.

the difficulty to maintain concentration and inhibit impulsive responses. Other neurocognitive pathologies, such as schizophrenia, have also been associated with detrimental effects on oculomotor control (e.g., saccadic suppression; Lencer et al., 2021; Broerse et al., 2001; smooth-pursuit abnormalities; Hutton & Kennard, 1998; see Wolf, Ueda & Hirano, 2021 for comprehensive review), coinciding with the disorganized thinking and executive dysfunctions typical for the disorder. In another field, patients with clinical psychological conditions like post-traumatic stress disorder (PTSD) have been shown to respond positively to so-called *eye-movement desensitization and reprocessing* (Shapiro, 1989; Shapiro & Solomon, 2017; Wilson et al., 2018) treatments, which reduce the severity of intrusive quasi-sensory experiences of trauma. During narration of traumatic episodes, the patient is instructed to keep gaze on rapidly moving visual percepts (e.g., the therapist's fingers), which helps to disrupt the vividness and momentary emotional impact of the traumatic memory, potentially by weakening the consolidation of distressing visual simulation as a conscious internal percept. Similar mechanisms may be at play during *rapid eye movement* (REM) sleep, where high-frequency eye movements accompany intense mental imagery and emotional memory processing, suggesting that even during sleep, eye movements are functionally related to dynamic perceptual simulations (Hong et al., 2018; Baird et al., 2022; see Mota-Rolim, 2020 for critical review).

### *Subchapter summary*

In summary, empirical findings from a variety of disciplines support the conclusion that spontaneous, non-visual eye movements are activated systematically by cognitive processes as well as the content of internal representations. Reduction of eye movement marks that attention is captured by something that is being internally generated or manipulated — unless this process is language-based, where keeping the eyes still motivates the mind to wander, yielding representations that would interfere with coherent message generation and speech planning. Eye movements indicate attention capture (rate) by simulations and focus on their content (travel distance), revealing a currently »epistemically privileged« (O'Callaghan, 2012: 88) knowledge representation in the cognitive tasks presented here.

#### 5.3.4.2. Horizontal bias in eye movements

Eye movements are seemingly reflexive, high-frequency bodily responses that may occur randomly, which casts doubt on the premise that eye movements are per se meaningful. They may be launched as a symptom of dry eyes, a bother of the eyelids, or other physiological incidents (like floaters). Without triangulation of eye data with concurrent neurological measurements or in situations without experimental control, investigations of non-visual gaze will always be limited by considerate noise.

One such type of noise might be an artefact of our evolutionary development. The present analysis found that most effects on travel distance occurred along the horizontal oculomotor axis. Across all participants, horizontal eye movements were consistently larger than vertical ones, suggesting a default bias of the oculomotor system toward launching saccades horizontally. Similar findings were reported before (Hansen & Essock, 2004: 1052; Demarais & Cohen, 1998: 236; Collewijn & Tamminga, 1984). Richardson and colleagues (2003: 775), for instance, reported that subjects' response times in a visual identification task were lower when the stimulus pictures were arranged horizontally as opposed to vertically. Danion and colleagues (2021) and Liu (2009) found that voluntary smooth pursuit of a vertically moving target was less consistent, i.e., moving fixations were held less precisely, than when the target moved horizontally. Eyeball rotation in vertical, horizontal and oblique direction is controlled by dedicated, separate cortical structures (Purves, 2012). Rotation amplitude of the eyeball in the horizontal axis is larger than in the vertical axis (Ohlendorf et al., 2022) and, in resting state, our eyes are open wider in horizontal diameter than vertical. It has been suggested that, as a species adapted to ground foraging, our visual system evolved to prioritize the detection of survival-relevant information primarily along the horizontal plane (Ryan & Shen, 2020). Moreover, as mentioned earlier, our ears are positioned on the sides of our heads, which makes us particularly adept at auditory localization along the horizontal axis relative to our body (McDermott, 2013).

Considering all these embodied constraints on eye movement from an enactivist point-of-view, human subjects might be biased to perceive motion as unfolding in the horizontal axis by default. This may affect how readily this study's participants recognized horizontal stimuli as instances of motion in contrast to events of vertical motion. Speculatively, horizontal motion might be overall more salient for human perceivers than vertical motion (cf. Footnote 90 above).

### 5.3.5. On the necessity of simulation for conceptual representation

Research that links blank screen eye movements to visual imagery faces an important, yet often unaddressed criticism. This criticism concerns the question whether simulations, here interpreted to manifest in eye movements, functionally contribute to conceptual representation, whether they occur after conceptual processing has finished, or whether they occur in parallel but without any functional interactions. This relates to the larger controversy touched upon in the introduction, namely whether the activation of representations from sensory modalities (perceptual simulation) is an automatic and necessary occurrence in cognition or whether these representations are co-activated after an abstract and inaccessible operation has produced a concept, based on which the sensory modalities are co-activated epiphenomenally. Empirical studies either do not take sides on this critical issue or make no mention of it, leaving their stance implicit in the conclusions they draw, which tend to lean more towards one or the other positions.

Accordingly, this subsection outlines main perspectives in this debate, evaluates whether the present findings can contribute to it, considers the suitability of eye-tracking as a methodology to provide answers, and suggests directions for future research.

#### 5.3.5.1. The ‘epiphenomenon criticism’ against 4E theories

Seminally, Mahon and Caramazza (2008) have challenged the view that modal representations are ubiquitous and causal in conceptual processing (Pulvermüller et al., 2005; Barsalou, 1999; 2008), arguing instead that concepts can be represented amodally and accessed independently of activity in perceptual systems. The authors cite neuropsychological and neuroimaging evidence showing that individuals with cortical lesions in sensorimotor regions can nevertheless perform conceptual tasks, suggesting that modality-specific activations are not strictly necessary for schematic inference or related deep processing (as posited by, e.g., Barsalou et al., 2008). Consequently, they propose that the sensorimotor activations frequently observed in conceptual tasks reflect epiphenomenal or context-dependent processes rather than the core format of conceptual representation (see also Pylyshyn, 1973; Fodor, 1975).

Applying this question of causal involvement to the present study of non-visual gaze: Are eye movements, as indicative of perceptual simulations, functionally involved

in conceptual representation, or do they occur after (or concurrently to) the fact and are triggered by what emerges from a more fundamental abstract and disembodied process?

#### 5.3.5.2. Competing views on eye movements and mental imagery

One proposal is related to enactivist theories from the 4E spectrum, essentially defending the claim that eye movements were necessary for phenomena like visual imagery to even arise. It is only the enactment of typical peripheral motor behavior or, rather, a mimicry of behavior reserved for exogenous perception, that enables such processes in high-level cognition. Mental imagery thus becomes quasi-sensory experience through enactment of typical behavior. This stance inherently assumes a feed-upward activation caused by activity of peripheral sensors (Spivey et al., 2000; Fuchs, 2018; Thomas, 1999). Cognition, in this perspective, is truly *embodied*: without oculomotor activity, there is no mental imagery.

Another widely held view is representationalist and sees eye movements arise as a concurrent response to emerging simulations. Experiencers of mental imagery execute eye movements because they scan a visual mental representation that has appeared to them and by doing so, they bring specific areas of the mental image into foveal focus or move the eyes along a previously encoded scanpath, again mimicking behavior of exogenous perception (Borst & Kosslyn, 2008). This view assumes a downward activation of the peripheral sensors parallel to higher-level activity, locating the original process in higher-level cognition and the brain.

These two stances differ strongly in the function they ascribe to eye movements. The enactivist position takes non-visual eye movements to be causal factors in the emergence of a conceptual representation, although many doubt that mentalizing in form of perceptual representations would be required (Thomas, 1999; Gibson, 1979; Neisser, 1976; Gallese, 2007; Fuchs, 2018)[96]. The representationalist position interprets non-visual eye movements during conceptual representation as responses to attentional shifts that occur during or after perceptual simulations provided the conceptual system with a processable representation (Kosslyn, 1994; Brandt & Stark, 1997; Laeng & Teodorescu, 2002; Borst & Kosslyn, 2012). While the enactivist view sees eye movements as integral to conceptualization, the representationalist view holds that eye movements only occur once conceptual processing has produced

[96] See Fodor & Pylyshyn (1988: 9ff.) for a brief discussion of eliminitavism vs. representationalism.

something comparable to a mental visual percept that yields oculomotor coordinates for saccade programming.

#### 5.3.5.3. What are the shortcomings of each view?

The following paragraphs sketch criticism of either view. First, the representationalist view is logically inconclusive in some respects. One of these open questions concerns the »feedback from the oculomotor representations to the imagery representations« (Spivey & Geng, 2001: 237). If eye movements are programmed based on visual mental representations, then what happens to these representations when the eyes move? Do they remain static in objective mental space, like when we observe a painting in a museum, or do they change position in our mental visual field? If they remain static, do they change in acuity because we shifted the location of the 'foveal' center? Findings from looking-at-nothing research (e.g., Laeng et al., 2014; Brandt & Stark, 1997) logically depend on the premise that this representation is static and retains the spatial relationships of the original stimulus (Kosslyn, Ball & Reiser, 1978). Hobson and colleagues (2014) propose that rapid eye movements during REM sleep reflect how the brain's predictive-coding networks are scanning dream imagery. Other blank screen studies examining imagery eye movements in long-term memory retrieval (Fourtassi et al., 2017; Johansson et al., 2018) suggest flexible morphing of visual representations, casting doubt on the notion that what we see with our mind's eye is static, let alone a mere copy of visual perception. Changing mental imagery by making eye movements would be identical to modifying the conceptual representation, attesting some functional influence of motor enactment and a tighter association of concept and percepts. Thus, the question whether a simulation-based conceptual representation is affected by eye movements or not lies at the heart of this debate. Examinations of simulation-based eye movements must deal with the question how stable the visual mental representation is that is used to program eye movements.

The very notion that experience of mental imagery was as if we were looking at images with our mind's eye is further criticized as a homonculus-problem. In neurocognitive terms: »As long as we do not know how the patterns of trees on our retinas evoke the concept of a "forest" in us, we also do not understand how reactivation of early visual brain areas can do this« (Hauk, 2016: 785). If eye movements were triggered based on information-processing of *internal* visual representations as if they were *external*, explanation is equally required for how we do

this internally with our mind's eye (cf. O'Regan & Noë, 2001). This shifts rather than solves the problem because it leaves unexplained how quasi-visual information is extracted and associated with concepts.

Taking another route, enactivists would argue that mental representation is not necessary, ergo, information-processing of internal visual stimuli is absent. The very motoric movement of the eyes is sufficient: »rather than being a spectator of a resemblance of [a perception, the individual] is resembling a spectator of [a perception, D.D.]« (Spivey et al., 2000: 6). But this seeming solution to the homonculus-problem cannot account for all findings either. With respect to the results of the present analysis, movement direction effects would be meaningless. Without a mental representation that biased saccades in specific directions, an analysis of direction effects on travel distances loses all explanatory value – especially for the environmental sounds, which were played in mono, rendering it impossible to hear how sound sources moved through space. Additionally, saccade rate would be expected to increase, because enacted processing of dynamic motion events in space would naturally increase eye movement activity. Movement direction did not have significant effects on rate in this study, however, contradicting that thinking about of movement space necessarily caused participants to make more eye movements.

Undeniable criticism of an enactive involvement of non-visual gaze in conceptualization comes from the vivid visual imagery that people commonly experience during reading (cf. Mar & Oatley, 2008). In terms of oculomotor behavior, reading requires high-frequency, mostly unidirectional saccadic shifts in the horizontal axis. Could such strongly increased, rhythmic eye movements be the cause for why we have strong visual imagery during reading? If so, what if the narrative prompts us to visualize something that does not lie in the horizontal plane — would that interfere? Would text-induced visuospatial imagery keep us from understanding the text, which appears to us as a visual percept itself (i.e., letters on paper)? Empirical findings do not license simple conclusions, although there are tendencies for increased blink rates and longer fixation times, i.e., reduction of oculomotor activity, during mindless reading versus attentive reading (Steindorf & Rummel, 2020; Foulsham et al., 2013; Smilek et al., 2010). Therefore, again, the crucial role that enactivist frameworks ascribe to eye movements for simulation-based conceptual processing cannot be confirmed. Eye movements cannot be a necessary condition for simulations.

#### 5.3.5.4. How future research could approach the epiphenomenon issue

Neuroscientific research has repeatedly found causal involvement of cortical structures that are responsible for sensory processing in tasks that require higher-level cognitive operations (Glenberg & Kaschak, 2002; Pulvermüller et al., 2005; Yee et al., 2013; Laeng & Teodorescu, 2002). Although these findings are criticized and debated (Johansson et al., 2012; Micic et al., 2010; Mahon & Caramazza, 2008), they do corroborate the claim that activation of perceptual symbols (Barsalou, 1999; 2009) for concept representation is empirically justified and not an unfounded theoretical hypothesis. Moreover, they make it hard to believe that cortical activity in modal brain regions is merely epiphenomenal to detached, higher-level mental operations. What these studies as well as their critics have failed to elucidate thus far, however, is the factors that control how conceptual processing becomes context-sensitive or task-dependent and, consequently, how strongly perceptual representations may play a role in what is termed amodal cognition (Mahon & Caramazza, 2008).

Additionally, previous studies have remained largely agnostic about the actual impact of perceptual simulations on subjects' conscious epistemic awareness, i.e., their experience of knowing. While it is unquestionable that people with cortical lesions suffer from deficiencies in daily life because physiological substrates of specific neural or motor functions are unavailable (i.e., losing abilities of space perception), healthy subjects may have developed, throughout their lives, individual preferences or neurocognitive styles to defocus or attenuate processing based on perceptual simulations. Depending on the severity of a lesion, clinical subjects may arrive at the same conclusions about common phenomena as healthy subjects and do not suffer from disadvantages in epistemic understanding. Furthermore, considering Barsalou's (2008) central argument that perceptual symbols are spread across the cortex, lesion-related deficiencies in a specific modality would not be expected to nivellate perceptual simulation altogether, because it is a domain-independent form of cognitive computation. Finally, as has been discussed above, sheer increased activity in the visual system during simulation is not equivalent to phenomenology of vivid visual imagery. While it has been shown that cortical excitability and activation strength is a correlate of the degree of post-hoc conscious imagery vividness, clarity on the predictive power of neurophysiologically measureable energy flow about subjective conscious experience remains an important desideratum (see Schwarzkopf, 2024).

#### 5.3.5.5. Augmenting the potential of eye-tracking

Eye-tracking methodology has large potential in the study of automatized, computational processes like perceptual simulation and is likely to enrich this debate in the future. It is a non-invasive method and has a high-enough temporal sampling frequency and spatial resolution to allow for detection of miniscule and short-lived systematic patterns in oculomotor behavior. Difficulties of such research endeavors arise from the transience of cognitive effects on non-visual gaze behavior — maybe it is only a couple of saccades or a single fixation that is triggered by the higher-level process under examination. Isolation of these individual oculomotor occurrences from the many others that coincide during, before, and after those in question must be based on empirical observation. Problems related to so-called *signal isolation* are common in neurocognitive experiments (e.g., motion artefacts in EEG).

A common resort for eye-tracking paradigms is to carefully look at theoretically motivated segments of time after stimulus onset. Barsalou and colleagues (2008: 274, citing Pulvermüller et al.'s (2005) word processing study) propose that as early as 200 ms post stimulus onset, simulation ensues to support conceptual processing. Hagoort and colleagues (2004: 440) examined semantic processing of language and concluded that meaning construction (N400) and syntactic integration (P600) are ongoing as early as 400 ms and 600 ms after stimulus onset. However, predictive processing in comprehension of natural speech or narratives strongly limits the use of such temporal anchors. Prediction allows us to anticipate potential prosodic, syntactic or semantic components of the continuous input, leading to pre-activation of potentially relevant concepts and anticipatory conceptualization, before they encounter the pertinent references in the input (Altmann & Kamide, 2004). Thus, relevant temporal segments of conceptualization of entities or relations might have already passed long before the surface form establishes concrete reference. Therefore, even in studies that aim to control for the content of simulations, time-locking of eye-tracking data to the unfolding verbal stimulus may not target the cognitive operation of interest.

Regardless, if eye movement patterns are detected around, e.g., a peak in the N400, and if they exhibit systematic characteristics across participants and tasks, a more data-driven conclusion about the signal character of eye movement patterns becomes possible. The additional data allows eye movements to be closely time-locked to measurements taken to be constitutive of conceptualization (i.e., causal) or considered a consequence of it (i.e., epiphenomenal). Causal involvement would entail

that simulation-related eye movement patterns occur before or during a conceptually driven decision or response, while epiphenomenal eye movements occur later.[97]

#### 5.3.5.6. The contribution of the present study

The central research question in this thesis was focused on the role of perceptual simulations in conceptual representation during language comprehension and language production. The main hypothesis tested whether systematic eye movements would mimic the extension of the moving figure's trajectory through motion event space. Detecting such eye movements would speak to visualizations of space as an integral part of the conceptual representation of motion.

Systematic eye movements did indeed show, and the statistical analysis confirmed an association of motion event space with these oculomotor responses. However, they fail to reveal anything about *when* in online cognition these eye movement-triggering sensory-based representations were co-activated and whether perceptual simulation helped establish the conceptual representation. The trial-wise aggregation into travel distances and rates does not allow for precise timing of those saccades that impacted travel distances so much in one direction or the other, making it impossible to classify them post-hoc as originating from an enactivist contribution to conceptual representation or to classify them as an epiphenomenal response to a visual representation that emerged from simulation. Therefore, the present analysis offers no answer as to whether simulations, manifesting in systematic eye movements, are functionally involved or epiphenomenal in conceptual representation. The present results can only provide speculative indications because the data was aggregated and thus temporally imprecise. First, perceptual simulation seems to have run concurrently to conceptual representation because the effects on aggregated eye movements occurred in the pertinent epochs and not later. Second, the degree to which perceptual simulations capture participants' attention seems to be dependent on the cognitive task

[97] Earlier plans of this projects' methodological and analytical procedure focused more strongly on considering the temporal dimension. An unpublished pilot study (n=80 participants) aimed to determine average recognition points in the environmental sound stimuli. Additionally, the original analysis plan was to model continuous gaze patterns with methods for time-course data. This plan was changed, first, to target individual saccades with precise onset timestamps, amplitudes, and directional angles, reducing the temporal dimension to individual onsets. Ultimately, aggregation of this saccade data into a time-independent variable was chosen, collapsing angles and amplitudes to two-dimensional travel distance. Reminiscent of the original, meticulous considerations is the data-driven epochal segmentation in Experiment 2, where participants' keypresses and voice onsets were taken as subject-and-trial-specific temporal anchors of online cognition.

(language vs. env. sound comprehension; language production vs. comprehension) and varies with participants' ability to experience vivid mental images.

Barsalou and colleagues' (2008) LASS Theory, which provided the foundational theoretical framework of this study, likewise assumes that conceptual representation is flexible and adaptable, and that different modes of processing may become active gradually (cf. also Barsalou, 2009). This is most certainly the case (cf. Truman & Kutas, 2024), but theoretical progress is not achieved with broad models that are not falsifiable. The occurrence of simulation-driven eye movements, as has been discussed at length here, is task-dependent, highly individualized, and seemingly unpredictable. While Barsalou (2008; et al., 2008; 2009; 2016) is correct to assume a flexible conceptual system (Kemmerer, 2015; Truman & Kutas, 2024), this characterization still leaves us far from knowing what exactly happens under the hood when we represent concepts. This experimental study was an attempt to move us one step closer.

#### 5.3.5.7. Summary

The present study took a small step towards resolving the debate whether our conceptual system crucially relies on co-activations of sensory-motor representations to achieve its functionality. The present results suggest, in line with Barsalou and colleagues' (2008) assumptions, that perceptual simulations run at least concurrently to processes of meaning construction, i.e., the representations activated in sensory modalities are activating *during* event model consolidation. However, aggregate travel distance is a variable largely dissociated from the temporal dimension, thus preventing insight into the precise time course of simulation-based eye movement patterns, rendering the results limited for discussions of causality. Future research on this issue will benefit from triangulation of gaze data with more direct, neural signatures of online cognition. While eye-tracking by itself cannot contribute any more to this discussion than other methods already have, this study shows that it still constitutes a fruitful method in research incentives that target the knotty entanglements of consciousness, attention, and mental representation.

### 5.3.6. Limitations

Multiple limitations of this study are acknowledged. Some are theoretical in nature, others refer to the analysis and data collection procedure.

#### 5.3.6.1. Limitations of the data collection

##### 5.3.6.1.1. Fixating a blank screen

The first limitation stems from the task design and technical peculiarities related to eye-tracking systems such as the one used in this study. For eye-tracking data to be largely noise-free and consistent, video-based eye-tracking seems to require that pupil size does not increase beyond ranges that are atypical with respect to participants' eye physique. Decrease of tracking quality due to pupil size fluctuations has been well-documented (Fink et al., 2024). Manuals therefore recommend that luminance to the eyes is kept constant and, due to pupil foreshortening (Brisson et al., 2013), that relevant visual stimuli are not shown in extreme positions of the trackable display area. Despite following these recommendations, gaze estimation accuracy was found to be impacted more negatively in participants whose pupils dilated so much that they compressed the iris to a minimum.

Specifically, when pupils were enlarging, the two gaze cursors, representing position of foveal gaze, slowly drifted apart and moved back closer together — as if each eye was independently in smooth pursuit. Considering that the gaze estimation method relies on the center-of-mass of a circular area drawn around the pupil, and considering that pupils are not perfectly circular, nor dilate in a perfectly circular fashion, this drifting apart of the gaze cursors may have been caused by irregular pupil dilation patterns in participants, causing temporary tracking errors. Since the only visual percept on the screen was the medium-grey square on an otherwise black background during recording, the extreme fluctuations of pupil size were unlikely caused by luminance fluctuations. Instead, they were likely caused by cognitive load (Kahneman & Beatty, 1966; Beatty, 1982) related to the memory task. Pupillometric research has shown that pupil size increases proportionately with the number of items that have to be processed in memory tasks (Wahn et al., 2018; van der Wel & van Steenbergen, 2018), easily reaching maximum dilation diameter in studies with multiple stimuli. The memory task here required participants to encode 12 items into

memory. This potentially induced significant cognitive load, which co-occurred with rapid phasic increases of pupil size to the detriment of tracking accuracy.

A related limitation may arise from the blank screen itself and the requirement to keep gaze within the grey square. To comply with this demand, some participants might have voluntarily stared at the same position in the square and, by focusing to not lose this control over oculomotor behavior, suppressed spontaneous eye movements. This staring may have introduced higher demands on their attentional capacities (Martarelli & Mast, 2013; but see Bochynska & Laeng, 2015), making them unavailable for focused stimulus processing. These participants would likely exhibit inferior recognition accuracy in the recall phase. However, since recall accuracy was not relevant in the present study, examination of this potential limitation will be a matter of further research.

Consequently, open questions remain about the use of blank screen memory tasks to capture the spontaneous, non-visual gaze behavior that characterizes cognitively driven oculomotor activity. Their chief strength is the absence of meaningful visual stimulation, yet the prolonged attentional demands of staring at an empty screen may themselves introduce confounds.

#### 5.3.6.1.2. Blockwise mental models

Another potential limitation may have arisen from participant strategies. Despite stimulus randomization, some participants mentioned that item ordering made the stimulus events appear as if they were part of some larger, global scene (e.g., typical events in an urban environment). Thus, in addition to conceptualizing the individual stimulus events, they configured the event models into a structure that connected them to something comparatively meaningful, as if they constructed mental models of entire encoding blocks as a parallel strategy to relieve memory load.

Since integration of individual events into such larger structures necessarily requires more abstract, schematic memory representation, loss of specific details possibly induced more recall errors and longer response times. One participant mentioned that they were not sure whether they had heard one or another sound during encoding, suggesting that they encoded a more abstract representation of the event and forgot about particulars (e.g., mistaking footsteps on marble vs. wood floor). Another doubted whether they had heard the ‘running faucet’ before because they were sure to have encoded ‘pouring water in a glass’, suggesting that they activated

attributes of a ‘water-pouring’ schema during encoding (or, e.g., ‘horse-neighing’ and ‘horse-trotting’). It remains a question of further research in what ways such higher-level structuring of event representations influences memory performance (see Gloede & Gregg, 2019; Röder & Rösler, 2003).

What this certainly suggests is that the event models retrieved during recall may not exhibit a similar conceptual make-up as was encoded. The recalled event models may be more schematic and coarser, representing rather the most frequently encountered or central conceptual features of the event schemata (Gerwien & von Stutterheim, 2018). In semantic memory, superordinate categories hold more abstract features and inherit these to subordinate concepts, although these inherited attributes tend to become less salient at lower levels (Anderson, 2013). However, this study cannot determine whether more abstract encoding strategies influenced participants’ spontaneous gaze behavior because use of these strategies was neither manipulated nor measured. Consequently, it was not treated as a systematic variable or examined as a dedicated experimental condition (e.g., abstract encoders vs. the rest).

#### 5.3.6.1.3. Perceiving a sound but encoding a word

Remembering events based on sound only is not an intuitive task we encounter day-to-day — usually vision and other senses converge to support our encoding and interpreting of sensory information for cognitive processing of our surroundings. It has been shown that auditory recognition memory is worse than visual recognition memory (Cohen et al., 2009; Gloede & Gregg, 2019), but this disparity can be relieved with training (e.g., musical expertise; Cohen et al., 2011). Considering that the environmental sound memory task essentially required auditory object encoding based on sound only, participants may have used other strategies to improve their performance. One potential strategy is to store event models with verbal labels (Lupyan, 2012) and rehearse the labels before recall. Single words may even serve as a memory pointer. If this was the case, a crucial consequence would be that participants may have relied less on situated simulation during recall and instead more on verbal labels.

A few results speak against such verbal encoding strategies. First, only four out of 42 participants responded that they imagined things in words rather than images or other sensory formats. While imagery is certainly not an all-or-nothing phenomenon and conceptual processing is rather fluid in its reliance on verbal vs. sensory-based

representations (Barsalou et al., 2008), most participants attested that they used mental imagery to some degree (see Table 4-7 in Chapter 4.4.7.2). Second, in both memory tasks (H1) and in the audio-replay phase of the verbalization task (H5), travel distances were affected by the movement direction conditions, suggesting that eye movements responded to motion space in event models constructed from both non-verbal and verbal input. If inner speech, i.e., covert language production, had been used as a mnemonic technique (Alderson-Day & Fernyhough, 2015), these movement direction effects would have been less likely, as is attested by the travel distance data from the pre-button/pre-voice epochs in the verbalization task (H5). Third, previous research and present results from the comparison of the non-verbal and verbal condition (H1) indicate that processing language input as opposed to visual imagery significantly drives up eye movement rates in non-visual gaze behavior (Ehrlichman & Barrett, 1983; Demarais & Cohen, 1998; Antrobus, 1973). Consequently, use of inner speech during the non-verbal memory task would have likely resulted in saccade rates comparable to those of the verbal memory task. In fact, it is possible that participants shifted from processing with situated simulations to verbal processing within the trial, expressing higher saccade rates later in the trial. Due to aggregation, the temporal unfolding of oculomotor patterns was not analyzed here, but future research might examine systematic clusterings of saccades in non-visual gaze studies as chronometric indices of shifts in representational medium.[98]

A similar limitation holds for the verbal memory task. When participants are presented with verbal stimuli, they may have dominantly encoded the linguistic surface forms as opposed to event models, based on the logic of encoding specificity (Kent & Lamberts, 2008). However, research on memory from written narrative comprehension suggests that even in recall tasks after short delays, participants' blank screen gaze responses were more likely associated with a situation model representation versus a text-based representation (Johansson, Oren & Holmqvist, 2018). Furthermore, unless tasks explicitly require encoding of surface forms (e.g., to deliver punchlines of jokes), surface information of both verbal and visual stimuli tends to fade more rapidly than central conceptual features (Gernsbacher, 1985; Kintsch et al., 1990; Radvansky & Zacks, 2014: 58f.).

[98] A rigid application of encoding specificity to the environmental sounds would entail that participants predominantly encoded non-verbal, auditory representations (into so-called echoic memory). Röder and Rösler (2003) demonstrated that recognition memory performance is reduced when relying on the encoding of transient and acoustically variable features of environmental sounds, compared to encoding that engages more stable and structured event representations.

#### 5.3.6.1.4. Measurability of mental imagery

The reported variability in mental imagery makes reliable correlation of simulation effects difficult. Not only is the experience of mental imagery highly subjective and not always conscious, but there are also no non-invasive methods to measure the intensity of the experience objectively. Even if one compared, e.g., the excitability or size of a participants' visual cortices, this would not reveal anything of substance about the concrete phenomenology. To capture the latter, numerous introspective questionnaires were designed (see Chapter 3.3.5.3.3). However, standardized surveys are limited in capturing vividness of experience objectively, as it is an inherently subjective phenomenon. Further, these surveys are based on introspection and, not only do people differ in metacognitive abilities, but they might generally be unaware »of which modality they are thinking in« (Johansson et al., 2006: 1067). This problem of measurability creates challenges for empirical research (O'Callaghan, 2012) and, at the same time, motivates incentives to find more objective measuring tools.

### 5.3.6.2. Limitations of the analysis procedure

#### 5.3.6.2.1. Saccade-less trials

As described in the analysis chapter, saccade-less trials were not analyzed because no travel distance could be computed. While this seems logical from a statistical rationale (zeros are problematic), it affects the interpretability of the data. A main argument in the previous discussion presumes a meaningful relationship of oculomotor activity to cognitive processing, and while statistical associations were found, the analysis completely neglected such zero-trials. However, solid argumentation for a graded expression of simulation in eye movement patterns must account also for the null cases. Simply put, participants who did not move their eyes in a trial still processed the trial.

Recall the finding that participants whose attention was more strongly drawn to situated simulations (high-visualizers) exhibited attenuated oculomotor activity when compared to (low-visualizers). Following this rationale, trials in which oculomotor activity was not measurable would have to be interpreted, too, namely that attention-capture by perceptual simulation was so strong that oculomotor activity was blocked completely. Hence, these zero-trials could be highly relevant and would be predicted to occur predominantly in participants who experienced vivid visual imagery. Since saccade-less trials were removed before travel distance calculation, however, this

remains a matter of future analyses and, consequently, a limitation to the present findings.

Note, however, that saccade-less trials in many cases may have been caused by recording issues, because subpar tracking accuracy paired with actual eye movements simply put out data that was not clean enough for saccade detection. Taking a technical perspective, saccade-less trials may indicate not only factual absence of eye movement, but possibly too much eye movement that was task-unrelated and, thus, constitute both technically and theoretically problematic data. Likely caused by low tracking accuracy or inconsistent participant behavior (e.g., saccading to off-screen locations with extreme gaze angles; head movement in the headrest messing momentarily with calibration; squinting as an embodied expression of recognition difficulty). From the technical side, this limitation could have been prevented by stricter participant moderation and more repeated validation of calibration model accuracy.

Despite accepting this limitation on the relationship between visualization intensity and oculomotor activity, does the available data (H2) not paint a telling image? Although the sample was relatively small and imbalanced in its distribution across the levels of visualization intensity (see Table 4-7 in Chapter 4.4.7.2), significant and marginally significant differences were reported. In the memory task, the travel distance data of 11 participants (visualization = 5) and, at the other end, 4 participants (visualization = 2) were compared to that of 20 participants (visualization = 4). Besides this imbalance, their differences in horizontal travel distance were significant, suggesting that visualization intensity may in fact have an impact on oculomotor patterns. In inferential statistics, however, imbalances in sample sizes and especially smaller sample sizes (<25) increase the risk of inference errors (Knudson & Lindsey, 2014). For example, it might be that either no effects, or much stronger effects would have shown, had more participants been equally distributed across the levels of visualization intensity.

Irrespective of this discussion, the finding that participants' meta-cognitive evaluations gained through post-hoc introspection about their own thinking could be related to spontaneous oculomotor activity is not only empirically supported (Johansson et al., 2011; Sima, 2014) but also raises important questions about the measurability of non-visual gaze behavior as a marker for cognitive activity (see Section 5.3.4).

#### 5.3.6.2.2. Standardizing trial durations per condition

As described in detail in Chapters 3.3.4.3 and 4.4.4, the stimuli in both conditions varied in duration, not only across modalities, but also across conditions within the same modality. This affected especially the environmental sound stimuli, where horizontal stimuli were longer than vertical stimuli and both were longer than control stimuli. To remedy this difference, trial durations were standardized by condition. In other words, all durations were normalized to have a mean of zero and a standard deviation of 1 within the respective condition, thus nivellating the differences between conditions. In statistical terms, then, the continuous predictor 'trial duration (z-scored)' was no longer correlated with the categorical predictor 'movement direction condition'. So-called multicollinearity of predictors limits the interpretability of model coefficients (Farrar & Glauber, 1967), therefore it was necessary to rule it out.

However, this step potentially introduced a considerable analytical limitation. By physical imperative, eye movements unfold in time. Although saccades are quick and the preparatory and executed motor sequence is almost unnoticeable for the subject, executing a single saccade requires approximately 100 ms for planning, 20-80 ms for execution depending on the length of the saccade, and some more time for post-saccadic adjustments (Land, 2019). Typically, inter-saccadic intervals exhibit an average range of 200-300 ms before the next saccade is executed, although this interval has been found to shorten linearly in multi-saccade sequences with every consecutive occurrence (Ghahghaei & Verghese, 2015, Exp. 3; Kelly et al., 2019). To be fair, these physiological benchmarks were gained exclusively from visually-guided eye-tracking research — the motor-programming and sequencing of non-visual saccades may be different (and an empirical desideratum), as they are not triggered by external stimulation. Still, the natural mechanics of the oculomotor system determine that the occurrence of saccades is dependent on time.

Based on these specifics, the critical analytical limitation is that the significant differences in travel distances between movement direction conditions may have come about only because participants had more time to execute more saccades and thus accumulate more travel distance in the respective conditions. To remedy this, statistical models included both the movement direction condition and the z-scored trial durations as predictors, and saccade rate as a slope in the by-subject random effect. Statistically, these models estimated the effects of the movement direction condition on travel distance while considering that, first, longer trial durations (within condition) may cause

larger travel distances (within condition) and, second, that higher rates proportionately increase travel distance (independent of condition).

Since rate represents a ratio, that is, the number of saccades per second, it is standardized by time and serves as an additional remedy of the potential effect of duration on travel distance. Saccade rate is the mathematical basis of travel distance, revealing whether, e.g., large travel distances came about through high frequency-low amplitude saccades or low frequency-high amplitude saccades. This allows disentanglement of the travel distance data from condition-specific duration differences: the inclusion of rate as a random slope leads the model to consider these travel distance increases under the constraint of the underlying number of saccades. Importantly, the full model on rate revealed that these longer durations did not increase saccade rates to such degrees that one could infer that larger travel distances only came about because participants had more time to make eye movements. While the full models on travel distance generally detected an overall linear increase of travel distance with trial duration, controlling for rate confirmed that condition-specific increases in travel distance persisted nonetheless and, thus, were less dependent on the time available to execute saccades but instead on increases in saccade amplitudes in the critical conditions. By controlling for both trial length and saccade rate, the models showed that motion event simulations biased oculomotor programming so that non-visual saccades became larger but not more frequent.

Altogether, the applied remedy seems to have worked: effects of movement direction condition were isolated from design confounds through elaborate statistical analysis. However, validity of findings in future studies that compare such time-sensitive data like saccades time-locked to naturalistic stimuli, be it non-robotic speech or environmental sounds, will benefit greatly from equal stimulus durations across all conditions.

#### 5.3.6.2.3. Perceptual simulation may be recruited for identification

When a participant did not recognize a stimulus, that trial was removed from both the memory task as well as the verbalization task because it could not be determined that they interpreted the stimulus as intended, barring any inference about conceptual representation. While the exclusion of these trials was necessary for analytical inference, potentially valuable eye movement data were discarded.

From the perspective of LASS Theory, participants who struggled to categorize the environmental sounds possibly used situated simulations to help their identification or disambiguation of the input. If the auditory input was not sufficient to narrow down the situation to a specific event schema, but there was a vague understanding of the dynamic components (e.g., identifying an event of falling but failing to identify the falling object), this narrowing-down might have been supported by perceptual simulations generating different potential scenarios in which the sound could occur, browsing through candidate event schemata that are compatible with these vaguely understood, fragmentary components. Therefore, simulation-based eye movements may have been observable especially in trials where event model construction was impeded.

Psycholinguistics and auditory neuroscience have long defended the benefits of multimodal integration for identification or disambiguation (McDermott, 2013: 163). While the phonemic restoration effect in speech recognition relies on the immediate phonemic and syntactic context, problems arising from the cocktail-party effect are usually solved or at least remedied by visual co-validation: hearing-impaired participants tend to look at interlocutors' mouths to disambiguate phonemes that are difficult to distinguish in multi-person conversations in noisy environments. In other situations, we may read a text by an author whose voice is familiar, and we hear their voice as inner speech (verbal auditory imagery). When hiking through nature, we may not see an animal that caused a rattling in the bushes, but we can still infer from analysis of the auditory sensory stream from where approximately the sound came and we can almost immediately infer whether it was a large or a small animal to assess potential danger (cross-modal inference). In other words, when the primary modality for processing a specific input is insufficient for identification of what we hear, activating information from other modalities seems to occur automatically for disambiguation and conceptual elaboration, likely driven by perceptual simulations.

# 6. Conclusion

The objective of the present study was to contribute to an open debate in cognitive science that revolves around the question whether in the comprehension or production of language — an abstract symbol system with dedicated processing mechanisms — co-activations of sensory knowledge are necessary and automatic. The concrete research question was whether conceptual representation was driven by perceptual simulations (Barsalou, 2008). The research question was operationalized on the basis of an extensive review of theoretical models and empirical findings, ranging from event cognition theory (Radvansky & Zacks, 2014), to motion event representation (Talmy, 2000b), and grounded cognition (Barsalou, 2008). It was assumed that in order to conceptually represent the meaning, participants would establish event models (Radvansky & Zacks, 2014). It was further assumed that they would rely on perceptual simulations (Barsalou, 2008) to represent the meaning captured by these event models.

To investigate this, two eye-tracking experiments were designed. 42 participants were recruited and completed a comprehension task (Exp. 1) and a language production task (Exp. 2). During the experiments, participants saw nothing but a blank screen while auditory stimuli (n=64), which depicted common situations in daily life (e.g., a car is driving past), were played to them via headphones. In Experiment 1, the stimuli were presented once as environmental sounds (non-verbal condition) and as spoken event descriptions (verbal condition), enabling the examination of comprehension processes for different types of input. Complementing Experiment 1, the purpose of Experiment 2 was to examine language production and participants were instructed to verbalize the environmental sound stimuli. This enabled not only a close examination of conceptualization and speech planning processes (Levelt, 1989), but also allowed for a comparison of cognitive processes between language production and language comprehension.

Stimuli of the critical condition referred to motion events in which movement direction of the figure was systematically manipulated. Motion event stimuli of the vertical condition presented situations in which the depicted movement is oriented along the vertical axis (e.g., a glass is falling to the floor), whereas horizontal motion event stimuli (e.g., a car is driving past) depict movement that typically unfolds in the horizontal plane. Control stimuli referred to events void of translatory motion (e.g., a

dog is barking). A thorough review of empirical evidence suggested that if participants relied on perceptual simulations for conceptual representation of these motion events, the spatial and temporal dynamics of their eye movements on blank screens would be systematically affected by the content of their stimulus-related perceptual simulations. Since motion event schemata provide information about the movement trajectory of the figure (Talmy, 2000), these perceptual simulations were assumed to activate a visuospatial representation of the figure's movement trajectory and that this would affect the dynamics of their non-visual gaze behavior.

The following hypothesis was formulated: participants' spontaneous eye movements exhibit larger spatial dispersion in the axes corresponding to the movement direction of the motion event's figure. Horizontal gaze dispersion is larger after horizontal motion events, and vertical gaze dispersion is larger after vertical motion events, compared to the control events in which spatial biasing of eye movements would be absent.

The statistical analysis revealed noteworthy findings. The main hypothesis was only partially confirmed due to absence of axis-specific effects across conditions. However, when participants constructed the meaning for motion events of the horizontal and vertical condition combined, both in the verbal and non-verbal condition in comprehension (Exp. 1), as well as in the during stimulus perception in language production (Exp. 2), their non-visual eye movements exhibited larger gaze dispersion in comparison to when they processed non-motion events. These findings indicate that perceptual simulation of spatial components of motion event schemata likely drove the representation of the meaning of these events, both in utterance production and comprehension.

When linguistic knowledge became more important to solve the experimental tasks, i.e., in verbal (compared to the non-verbal) comprehension and in speech-planning (compared to stimulus perception) for production, oculomotor activity was systematically affected in another dimension. Saccade rate was higher and travel distance lower during verbal vs. non-verbal comprehension (Exp. 1). Similarly, saccade rate was higher in speech-planning epochs during language production (Exp. 2). Thus, it seems that when cognitive resources are used for processing linguistic representations, be that during grammatical encoding in production or during parsing in comprehension, oculomotor activity changes systematically. This finding was

interpreted as indicative of changes in cognitive resource allocation, namely that processing of linguistic representations (phonological, morphological, syntactic) demands specialized capacities, causing resources to be drawn away from perceptual simulation.

Bearing methodological implications, an unexpected finding emerged. Participants' self-reported visualization intensity was significantly associated with their gaze dispersion and, in some contexts, with saccade rate. Participants who experienced vivid visual imagery during the experiments exhibited smaller gaze dispersion, while those reporting weaker visual imagery moved their eyes further. Thus, participants seem to vary in how attentional resources are drawn to perceptual simulations and how easily their oculomotor system decouples from exogenous perception, which may affect their ability for focused processing of thus generated representations. Opposing assumptions of 4E cognition, not every participant comes equipped with equal disposition for simulation.

The discussion concerned alternative explanations of systematic non-visual gaze behavior and limitations, such as the difficulty to control what people simulate as well as the empirical variability in mental imagery ability. Crucially, a variety of factors may contribute to systematicity in eye movements, ranging from attention and cognitive load to psychopathology and motoric defaults of the eye movement system.

The results of this study contradict major assumptions of 4E cognition theories which argue that activity in sensorimotor modalities is automatic and necessary for any type of cognitive activity. At the same time, the findings underline that language processing is not encapsulated from sensory knowledge. The fact that perceptual simulation did not affect oculomotor activity in all conditions equally suggests that the construction of conceptual representations works differently under different circumstances — and while perceptual simulation is certainly a promising principle, it is neither ubiquitous, nor necessary every time we comprehend or produce language, especially during speech planning or syntactic parsing. Instead, representations generated by perceptual simulation seem to be drawn upon flexibly across different types of cognitive processes or representations. Research has only begun to understand the many intricacies of this computationally powerful conceptual system of ours — and while the present study might not constitute a large forward leap, it remains up to future research whether it will be perceived as moving in the right direction.

# References


Abbassi, E., Blanchette, I., Ansaldo, A. I., Ghassemzadeh, H., & Joanette, Y. (2015). Emotional words can be embodied or disembodied: the role of superficial vs. deep types of processing. *Frontiers in psychology*, *6*, 975.

Agus, T. R., Thorpe, S. J., Suied, C., & Pressnitzer, D. (2010). Characteristics of human voice processing. In *2010 IEEE International Symposium on Circuits and Systems (ISCAS)*, 509-512.

Alain, C., & Arnott, S. (2000). Selectively attending to auditory objects. *Frontiers in Bioscience, 5*(1), D202-D212.

Alain, C., Arnott, S., & Picton, T. (2001). Bottom–up and top–down influences on auditory scene analysis: Evidence from event-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(5), 1072-1089.

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & De Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, *23*(15), 1427-1431.

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin, 141*(5), 931-965.

Allopenna, P., 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.

Allport, D. A. (1985). Distributed memory, modular subsystems, and dysphasia. In Newman, S., & Epstein, R. (eds.). *Current perspectives in dysphasia*. New York: Churchill Livingstone, 32-60.

Altmann, G. (2004). Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm'. *Cognition*, *93*(2), B79-B87.

Altmann, G. (2011). The mediation of eye movements by spoken language. In Liversedge, S., Gilchrist, I., & Everling, S. (eds.). *The Oxford handbook of eye movements*. Oxford: University Press, 979-1003.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247-264.

Altmann, G., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of memory and language, 57,* 502-518.

Altmann, G., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive science*, *33*(4), 583-609.

Amsel, B. D., Urbach, T. P., & Kutas, M. (2014). Empirically grounding grounded cognition: The case of color. *Neuroimage*, *99*, 149-157.

Anderson, J. R. (2013). *Kognitive Psychologie*. (J. Funke, ed.) (7th ed.). Heidelberg: Springer.

Andrade, J., May, J., Deeprose, C., Baugh, S. J., & Ganis, G. (2014). Assessing vividness of mental imagery: The Plymouth sensory imagery questionnaire. *British journal of psychology*, *105*(4), 547-563.

Antrobus, J. S. (1973). Eye movements and nonvisual cognitive tasks. In Zikmund, V. (ed.) *The oculomotor system and brain functions*. London: Butterworth, 354-368.

Antrobus, J. S., Antrobus, J. S., & Singer, J. L. (1964). Eye movements accompanying daydreaming, visual imagery, and thought suppression. *The Journal of Abnormal and Social Psychology, 69*(3), 244-252.

Audacity Team (2014). Audacity®: Free Audio Editor and Recorder [Computer program]. Version 2.3.3 https://www.audacityteam.org/

Awh, E., Armstrong, K. M., & Moore, T. (2006). Visual and oculomotor selection: links, causes and implications for spatial attention. *Trends in cognitive sciences*, *10*(3), 124-130.

Baird, B., Tononi, G., & LaBerge, S. (2022). Lucid dreaming occurs in activated rapid eye movement sleep, not a mixture of sleep and wakefulness. *Sleep, 45*(4), zsab294.

Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of experimental psychology: human perception and performance, 19*(2), 250-267.

Ballas, J. A., & Howard, J. (1987). Interpreting the language of environmental sounds. *Environment and behavior*, *19*(1), 91-114.

Ballas, J. A., & Mullins, T. (1991). Effects of Context on the Identification of Everyday Sounds. *Human Performance, 4*(3), 199-219.

Ballas, J. A., & Sliwinski, M. J. (1986). Causal uncertainty in the identification of environmental sounds. *Technical Report ONR-86-1*. Washington, DC: Georgetown University.

Ballas, J. A., Sliwinski, M. J., & Harding, J. P. (1986). Uncertainty and response time in identifying nonspeech sounds. *Journal of the Acoustical Society of America, 79*, S47.

Bandler, R., & Grinder, J. (1979). *Frogs into Princes: Neuro Linguistic Programming*. Moab, UT: Real People Press.
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. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577-660.
Barsalou, L. (2008). Grounded Cognition. *Annual Review of Psychology, 59,* 617-645.
Barsalou, L. (2009). Simulation, situated conceptualization, and prediction. *Philosophical transactions of The Royal Society B: biological sciences*, *364*(1521), 1281-1289.
Barsalou, L. (2016). Situated conceptualization: Theory and applications. In Coello, Y., & Fischer, M. (eds.). *Foundations of embodied cognition: Perceptual and emotional embodiment*. London: Routledge/Taylor & Francis Group, 11-37.
Barsalou, L. (2017). Cognitively Plausible Theories of Concept Composition. In Hampton, J., & Winter, Y. (eds). *Compositionality and Concepts in Linguistics and Psychology. Language, Cognition, and Mind, vol 3*. Cham: Springer, 9-30.
Barsalou, L. (2021). Categories at the interface of cognition and action. In Mauri, C., Fiorentini, I., & Goria, E. (eds.). *Building categories in interaction: Linguistic resources at work*. Amsterdam: John Benjamins, 35-72.
Barsalou, L., 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.
Barsalou, L., Santos, A., Simmons, W., & Wilson, C. (2008). Language and simulation in conceptual processing. In de Vega, M., Glenberg, A., & Graesser, A. (eds.). *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford: University Press, 245-283.
Bartlett, F. C. (1932/1995). *Remembering: A study in experimental and social psychology*. Cambridge: University Press.
Bartolomeo, P. (2008). The neural correlates of visual mental imagery: An ongoing debate. *Cortex*, *44*(2), 107-108.
Barwise, J., & Perry, J. (1983). *Situations and attitudes*. Cambridge MA: MIT Press.
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48.
Bates, E., D'Amico, S., Jacobsen, T., et al. (2003). Timed picture naming in seven languages. *Psychonomic bulletin & review*, *10*, 344-380.
Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276–292.
Bergen, B. K., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive science*, *31*(5), 733-764.
Bergen, B., & Wheeler, K. (2010). Grammatical aspect and mental simulation. *Brain and language*, *112*(3), 150-158.
Berman, R., & Slobin, D. I. (1994). *Relating Events in Narrative: A Crosslinguistic Developmental Study.* Hillsdale, N.J.: Lawrence Erlbaum Associates.
Bey, C., & McAdams, S. (2002). Schema-based processing in auditory scene analysis. *Perception & psychophysics*, *64*, 844-854.
Bierwisch, M., & Schreuder, R. (1992). From concepts to lexical items. *Cognition, 42*(1-3), 23-60.
Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, *14*(10), 693-707.
Blazhenkova, O., & Kozhevnikov, M. (2009). The new object-spatial-verbal cognitive style model: Theory and measurement. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *23*(5), 638-663.
Blazhenkova, O., & Kozhevnikov, M. (2010). Visual-object ability: A new dimension of non-verbal intelligence. *Cognition*, *117*(3), 276-301.
Bocanegra, B. R., Poletiek, F. H., & Zwaan, R. A. (2022). Language concatenates perceptual features into representations during comprehension. *Journal of memory and language*, *127*, 104355.
Bochynska, A., & Laeng, B. (2015). Tracking down the path of memory: eye scanpaths facilitate retrieval of visuospatial information. *Cognitive processing*, *16*, 159-163.
Bohnemeyer, J., & Pederson, E. (eds.). (2010). *Event representation in language and cognition*. Cambridge: University Press.
Bonner, M. F., & Grossman, M. (2012). Gray matter density of auditory association cortex relates to knowledge of sound concepts in primary progressive aphasia. *Journal of Neuroscience*, *32*(23), 7986-7991.
Borst, G., & Kosslyn, S. M. (2008). Visual mental imagery and visual perception: Structural equivalence revealed by scanning processes. *Memory & cognition*, *36*(4), 849-862.

Borst, G., & Kosslyn, S. M. (2012). Scanning visual mental images: Some structural implications, revisited. In Gyselinck, V., & Pazzaglia, F. (eds.). *From mental imagery to spatial cognition and language*. London: Psychology Press, 19-42.
Bower, G. H., & Morrow, D. G. (1990). Mental models in narrative comprehension. *Science*, *247*(4938), 44-48.
Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of cognitive neuroscience*, *9*(1), 27-38.
Bregman, A. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge MA: MIT Press.
Brehm, L., & Alday, P. M. (2022). Contrast coding choices in a decade of mixed models. *Journal of memory and language*, *125*, 104334.
Brewer, W. F., & Nakamura, G. V. (1984). The nature and functions of schemas. *Center for the Study of Reading Technical Report; no. 325*.
Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavior research methods*, *45*, 1322-1331.
Brockhoff, L., Elias, E. A., Bruchmann, M., Schindler, S., Moeck, R., & Straube, T. (2023). The effects of visual perceptual load on detection performance and event-related potentials to auditory stimuli. *NeuroImage*, *273*, 120080.
Broerse, A., Crawford, T. J., & den Boer, J. A. (2001). Parsing cognition in schizophrenia using saccadic eye movements: a selective overview. *Neuropsychologia*, *39*(7), 742-756.
Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., & Rizzolatti, G. (2005). Listening to action-related sentences modulates the activity of the motor system: a combined TMS and behavioral study. *Cognitive Brain Research*, *24*(3), 355-363.
Carlson-Radvansky, L. A., Covey, E. S., & Lattanzi, K. M. (1999). "What" Effects on "Where": Functional Influences on Spatial Relations. *Psychological Science*, *10*(6), 516-521.
Carlson-Radvansky, L. A., & Jiang, Y. (1998). Inhibition Accompanies Reference-Frame Selection. *Psychological Science*, *9*(5), 386-391.
Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance, 27*(1), 115-127.
Carroll, M., Weimar, K., Flecken, M., Lambert, M., & Stutterheim, C. von (2012). Tracing trajectories: Motion event construal by advanced L2 French-English and L2 French-German speakers. *Language, Interaction and Acquisition*, *3*(2), 202-230.
Chafe, W. (1970). *Meaning and the Structure of Language*. University of Chicago Press.
Charlot, V., Tzourio, N., Zilbovicius, M., Mazoyer, B., & Denis, M. (1992). Different mental imagery abilities result in different regional cerebral blood flow activation patterns during cognitive tasks. *Neuropsychologia*, *30*(6), 565-580.
Chatterjee, A., Southwood, M. H., & Basilico, D. (1999). Verbs, events and spatial representations. *Neuropsychologia*, *37*(4), 395-402.
Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge MA: MIT Press.
Cohen, M. A., Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). Auditory and visual memory in musicians and nonmusicians. *Psychonomic bulletin & review*, *18*, 586-591.
Cohen, M. A., Horowitz, T. S., & Wolfe, J. M. (2009). Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences*, *106*(14), 6008-6010.
Colby, C. L., & Goldberg, M. E. (1999). Space and attention in parietal cortex. *Annual review of neuroscience*, *22*(1), 319-349.
Collewijn, H., & Tamminga, E. P. (1984). Human smooth and saccadic eye movements during voluntary pursuit of different target motions on different backgrounds. *The Journal of physiology*, *351*(1), 217-250.
Craver-Lemley, C., & Arterberry, M. E. (2001). Imagery-induced interference on a visual detection task. *Spatial vision*, *14*(2), 101-119.
Craver-Lemley, C., & Reeves, A. (1992). How visual imagery interferes with vision. *Psychological review*, *99*(4), 633-649.
Cusack, R., Decks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis. *Journal of Experimental Psychology: Human Perception and Performance, 30*(4), 643-656.
Daimon, T. (2010). Box-Cox transformation. In Lovric, M. (ed.). *International Encyclopedia of Statistical Science*. New York: Springer, 177.
Dance, C. J., Ward, J., & Simner, J. (2021). What is the Link Between Mental Imagery and Sensory Sensitivity? Insights from Aphantasia. *Perception*, *50*(9), 757-782.

Danion, F. R., Mathew, J., Gouirand, N., & Brenner, E. (2021). More precise tracking of horizontal than vertical target motion with both the eyes and hand. *Cortex*, *134*, 30-42.
Darwin, C. & Carlyon, R. (1995). Auditory Grouping. In Moore, B. (ed.). *Hearing. Handbook of Perception and Cognition*. New York: Academic Press, 387-424.
Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, *23*(8), 3423-3431.
Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and cognition*, *135*, 103569.
Demarais, A. M., & Cohen, B. H. (1998). Evidence for image-scanning eye movements during transitive inference. *Biological psychology*, *49*(3), 229-247.
Deubel, H., Schneider, W. X., & Paprotta, I. (1998). Selective dorsal and ventral processing: Evidence for a common attentional mechanism in reaching and perception. *Visual cognition*, *5*(1-2), 81-107.
van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic bulletin & review*, *25*, 2005-2015.
Diamantopoulos, G., Woolley, S., & Spann, M. (2009). A Critical Review of Past Research into the Neuro-Linguistic Programming Eye - Accessing Cues Model. *Current Research in NLP*, *1*, 8-22.
Dick, F., Krishnan, S., Leech, R., & Saygin, A. P. (2016). Environmental sounds. In Hickok, G., & Small, S. (eds). *Neurobiology of language*. New York: Academic Press, 1121-1138.
van Dijk, T., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.
Dijkstra, K., & Post, L. (2015). Mechanisms of embodiment. *Frontiers in psychology*, *6*, 1525.
Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in cognitive sciences*, *23*(5), 423-434.
Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, *110*(3), 412-431.
Dove, G. (2022). Rethinking the role of language in embodied cognition. *Phil. Trans. R. Soc. B*, *378*: 20210375.
Dove, G., Barca, L., Tummolini, L., & Borghi, A. (2022). Words have a weight: language as a source of inner grounding and flexibility in abstract concepts. *Psychological research*, *86*, 2451-2467.
Dronkers, N. F., Wilkins, D. P., Van Valin Jr, R. D., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, *92*(1-2), 145-177.
Dudschig, C., Souman, J., Lachmair, M., Vega, I. D. L., & Kaup, B. (2013). Reading "sun" and looking up: The influence of language on saccadic eye movements in the vertical dimension. *PloS one*, *8*(2), e56872.
Dully, J., McGovern, D. P., & O'Connell, R. G. (2018). The impact of natural aging on computational and neural indices of perceptual decision making: A review. *Behavioural brain research*, *355*, 48-55.
Eckardt, B. von (2012). The representational theory of mind. In Frankish, K., & Ramsey, W. (eds.). *Cambridge handbook of cognitive science*. Cambridge: University Press, 29-50.
Ehrlichman, H., & Barrett, J. (1983). 'Random' saccadic eye movements during verbal-linguistic and visual-imaginal tasks. *Acta psychologica*, *53*(1), 9-26.
Ehrlichman, H., & Micic, D. (2012). Why do people move their eyes when they think? *Current Directions in Psychological Science*, *21*(2), 96-100.
Eisenberg, M., Zacks, J. & Flores, S. (2018). Dynamic prediction during perception everyday events. *Cognitive Research: Principles and Implications, 3*(53), 1-12.
Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision research*, *43*(9), 1035-1045.
Engel, K. C., Flanders, M., & Soechting, J. F. (2002). Oculocentric frames of reference for limb movement. *Archives italiennes de biologie*, *140*(3), 211-219.
Estes, Z., Verges, M., & Barsalou, L. W. (2008). Head up, foot down: Object words orient attention to the objects' typical location. *Psychological Science*, *19*(2), 93-97.
Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*, *49*(1), 92-107.
Ferreira, F., Bailey, K., & Ferraro, V. (2002). Good-Enough Representations in Language Comprehension. *Current Directions in Psychological Science*, *11*(1), 11-15.
Fink, L., Simola, J., Tavano, A., Lange, E., Wallot, S., & Laeng, B. (2024). From pre-processing to advanced dynamic modeling of pupil data. *Behavior Research Methods*, *56*(3), 1376-1412.
Finke, R. A. (1989). *Principles of mental imagery.* Cambridge MA: MIT Press.
Fillmore, C. (1968). The Case for Case. In Bach, E., & Harms, R. (eds). *Universals in Linguistic Theory*. London: Holt, Rinehart and Winston, 1-25.
Flecken, M., Athanasopoulos, P., Kuipers, J. R. & Thierry, G. (2015). On the road to somewhere: Brain potentials reflect language effects on motion perception. *Cognition, 141,* 41-51.

Flecken, M., Carroll, M., Weimar, K., & Stutterheim, C. von (2015). Driving along the road or heading for the village? Conceptual differences underlying motion event encoding in French, German, and French–German L2 users. *The Modern Language Journal*, *99*(S1), 100-122.
Flecken, M., Stutterheim, C. von, & Carroll, M. (2014). Grammatical aspect influences motion event perception: findings from a cross-linguistic non-verbal recognition task. *Language and Cognition, 6,* 45-78.
Floridou, G. A., Peerdeman, K. J., & Schaefer, R. S. (2022). Individual differences in mental imagery in different modalities and levels of intentionality. *Memory & cognition*, *50*(1), 29-44.
Fodor, J. (1975). *The language of thought*. Cambridge MA: Harvard University Press.
Fodor, J. (1983). *Representations: Philosophical essays on the foundations of cognitive science*. Cambridge: MIT Press.
Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3-71.
Foroni, F. (2015). Do we embody second language? Evidence for 'partial' simulation during processing of a second language. *Brain and Cognition*, *99*, 8-16.
Foroni, F., & Semin, G. (2013). Comprehension of action negation involves inhibitory simulation. *Frontiers in Human Neuroscience*, *7*, 209.
Foulsham, T., Farley, J., & Kingstone, A. (2013). Mind wandering in sentence reading: Decoupling the link between mind and eye. *Canadian Journal of Experimental Psychology, 67*(1), 51-59.
Fourtassi, M., Rode, G. & Pisella, L. (2017). Using eye movements to explore mental representations of space. *Annals of Physical and Rehabilitation Medicine, 60*, 160-163.
Frances, C. (2024). Good enough processing: what have we learned in the 20 years since Ferreira et al. (2002)? *Frontiers in psychology*, *15*, 1323700.
Freyd, J. J., & Finke, R. A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 126-132.
Fried, M., Tsitsiashvili, E., Bonneh, Y. S., Sterkin, A., Wygnanski-Jaffe, T., Epstein, T., & Polat, U. (2014). ADHD subjects fail to suppress eye blinks and microsaccades while anticipating visual stimuli but recover with medication. *Vision research*, *101*, 62-72.
Friederici, A. D., Gunter, T. C., Hahne, A., & Mauth, K. (2004). The relative timing of syntactic and semantic processes in sentence comprehension. *NeuroReport*, *15*(1), 165-169.
Friedman, N. P., & Robbins, T. W. (2022). The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, *47*(1), 72-89.
Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*, *62*(2), 1230-1233.
Fuchs, T. (2018). *Ecology of the brain: The phenomenology and biology of the embodied mind*. Oxford: University Press.
Gallese, V. (2007). Before and below 'theory of mind': embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 659-669.
Gallese, V., & Lakoff, G. (2005). The Brain's concepts: the role of the Sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, *22*(3-4), 455-479.
Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive psychology*, *17*(3), 324-363.
Gerwien, J., Filip, M., & Smolík, F. (2024). Noun imageability and the processing of sensory-based information. *Quarterly Journal of Experimental Psychology*, *77*(10), 2137-2150.
Gerwien, J. & Stutterheim, C. von (2018). Event segmentation. Cross-linguistic differences in verbal and non-verbal tasks. *Cognition, 180,* 225-237.
Gerwien, J., & Stutterheim, C. von (2022). Conceptual blending across ontological domains—References to time and space in motion events by Tunisian Arabic speakers of L2 German. *Frontiers in Communication*, *7*, 856805.
Gerwien, J., Stutterheim, C. von, & Rummel, J. (2022). What is the interference in "verbal interference"? *Acta Psychologica*, *230*, 103774.
Ghahghaei, S., & Verghese, P. (2015). Efficient saccade planning requires time and clear choices. *Vision research*, *113*, 125-136.
Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston MA: Houghton Mifflin.
Gleitman, L. & Papafragou, A. (2012). Language and Thought. In Holyoak, K. & Morrison, R. (eds.). *Cambridge Handbook of Thinking and Reasoning*. Cambridge: University Press, 633-661.
Glenberg, A. & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review, 9*(3), 558-565.
Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, *43*(3), 379-401.

Gloede, M. E., & Gregg, M. K. (2019). The fidelity of visual and auditory memory. *Psychonomic Bulletin & Review*, *26*, 1325-1332.
Gold, J. I., & Shadlen, M. N. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature*, *404*(6776), 390-394.
Greenberg, J. H. (1966). *Language universals*. The Hague: Mouton.
Grice, H. P. (1975). "Logic and Conversation". In Kimball, J. (ed.). *Syntax and Semantics, Vol. 3*. Leiden: Brill, 41-58.
Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological science*, *11*(4), 274-279.
Griffiths, T., & Warren, J. (2004). What is an auditory object? *Nature Reviews Neuroscience 5*, 887-892.
Gurtner, L. M., Hartmann, M., & Mast, F. W. (2021). Eye movements during visual imagery and perception show spatial correspondence but have unique temporal signatures. *Cognition*, *210*, 104597.
Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M., & Oxenham, A. J. (2005). Neuromagnetic correlates of streaming in human auditory cortex. *Journal of Neuroscience*, *25*(22), 5382-5388.
Gygi, B., Kidd, G. R., & Watson, C. S. (2004). Spectral-temporal factors in the identification of environmental sounds. *The Journal of the Acoustical Society of America*, *115*(3), 1252-1265.
Habel, C., & Stutterheim, C. von (eds.) (2000). *Räumliche Konzepte und sprachliche Strukturen*. Tübingen: Niemeyer.
Hagoort, P. (2017). The core and beyond in the language-ready brain. *Neuroscience & Biobehavioral Reviews*, *81*, 194-204.
Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*, 438-441.
Handjaras, G., Ricciardi, E., Leo, A., et al. (2016). How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *Neuroimage*, *135*, 232-242.
Hanning, N. M., & Deubel, H. (2018). Independent effects of eye and hand movements on visual working memory. *Frontiers in Systems Neuroscience*, *12*, 37.
Hansen, B. C., & Essock, E. A. (2004). A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. *Journal of vision*, *4*(12), 1044-1060.
Harding, S., Cooke, M., & König, P. (2007). Auditory Gist Perception: An Alternative to Attentional Selection of Auditory Streams? In Paletta, L., & Rome, E. (eds). *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*. WAPCV 2007. Lecture Notes in Computer Science, vol 4840. Heidelberg: Springer, 399-416.
Harnad, S. (1990). The symbol grounding problem. *Physica D: nonlinear phenomena, 42*(1-3), 335-346.
Hartmann, M., Mast, F. W., & Fischer, M. H. (2015). Spatial biases during mental arithmetic: Evidence from eye movements on a blank screen. *Frontiers in psychology*, *6*, 12.
Hauk, O. (2016). What does it mean? A review of the neuroscientific evidence for embodied lexical semantics. In Hickok, G., & Small, S. (eds.). *Neurobiology of language*. New York: Academic Press, 777-788.
Hebb, D. O. (1968). Concerning imagery. *Psychological review*, *75*(6), 466-477.
Herff, S. A., McConnell, S., Ji, J. L., & Prince, J. B. (2022). Eye Closure Interacts with Music to Influence Vividness and Content of Directed Imagery. *Music & Science*, *5*.
Herskovits, A. (1986). *Language and spatial cognition*. Cambridge: University Press.
Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in cognitive sciences*, *4*(4), 131-138.
Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience*, *8*(5), 393-402.
Hobson, J. A., Hong, C., & Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Frontiers in psychology*, *5*, 1133.
Hochberg, J., & Fallon, P. (1976). Perceptual analysis of moving patterns. *Science*, *194*(4269), 1081-1083.
Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & psychophysics*, *57*(6), 787-795.
Hong, C., Fallon, J. H., Friston, K. J., & Harris, J. C. (2018). Rapid eye movements in sleep furnish a unique probe into consciousness. *Frontiers in psychology*, *9*, 2087.
Huber, S., & Krist, H. (2004). When is the ball going to hit the ground? Duration estimates, eye movements, and mental imagery of object motion. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 431-444.

Huette, S., Winter, B., Matlock, T., & Spivey, M. (2012). Processing motion implied in language: eye-movement differences during aspect comprehension. *Cognitive Processing*, *13*, 193-197.
Humphries, C., Willard, K., Buchsbaum, B., & Hickok, G. (2001). Role of anterior temporal cortex in auditory sentence comprehension: an fMRI study. *Neuroreport*, *12*(8), 1749-1752.
Hutchinson, J. B., & Barrett, L. F. (2019). The Power of Predictions: An Emerging Paradigm for Psychological Research. *Current Directions in Psychological Science*, *28*(3), 280-291.
Hutton, S., & Kennard, C. (1998). Oculomotor abnormalities in schizophrenia: a critical review. *Neurology*, *50*(3), 604-609.
Indefrey, P. (2011). The spatial and temporal signatures of word production components: a critical update. *Frontiers in psychology*, *2*, 255.
Ishai, A., Ungerleider, L. G., & Haxby, J. V. (2000). Distributed neural systems for the generation of visual images. *Neuron*, *28*(3), 979-990.
Ishai, A., & Sagi, D. (1995). Common mechanisms of visual imagery and perception. *Science*, *268*(5218), 1772-1774.
Isnard, V., Chastres, V., Viaud-Delmon, I., & Suied, C. (2019). The time course of auditory recognition measured with rapid sequences of short natural sounds. *Scientific reports*, *9*(1), 8005.
Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, *40*(10-12), 1489-1506.
Iverson, P. (1995). Auditory stream segregation by musical timbre: effects of static and dynamic acoustic attributes. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(4), 751-763.
Jackendoff, R. (1996). Conceptual semantics and cognitive linguistics. *Cognitive Linguistics, 7*(1), 93-129.
Jackendoff, R. (2002). *Foundations of Language*. Oxford: University Press.
Johansson, R. (2013). *Tracking the mind's eye: eye movements during mental imagery and memory retrieval.* PhD thesis (Lund University).
Johansson, R., Holsanova, J., Dewhurst, R., & Holmqvist, K. (2012). Eye movements during scene recollection have a functional role, but they are not reinstatements of those produced during encoding. *Journal of Experimental Psychology: Human Perception and Performance, 38*(5), 1289–1314.
Johansson, R., Holsanova, J. & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive science, 30,* 1053-1079.
Johansson, R., Holsanova, J., & Holmqvist, K. (2011). The dispersion of eye movements during visual imagery is related to individual differences in spatial imagery ability. In *Proceedings of the annual meeting of the cognitive science society, 33*(33).
Johansson, R., Oren, F., & Holmqvist, K. (2018). Gaze patterns reveal how situation models and text representations contribute to episodic text memory. *Cognition*, *175*, 53-68.
Johnson, M. (1987). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago: University of Chicago Press.
Johnson-Laird, P. (1980). Mental models in cognitive science. *Cognitive science*, *4*(1), 71-115.
Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, *87*(4), 329-354.
Just, M. A., Newman, S. D., Keller, T. A., McEleney, A., & Carpenter, P. A. (2004). Imagery in sentence comprehension: an fMRI study. *Neuroimage*, *21*(1), 112-124.
Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583-1585.
Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of memory and language*, *49*(1), 133-156.
Kamide, Y., Lindsay, S., Scheepers, C., & Kukona, A. (2016). Event processing in the visual world: Projected motion paths during spoken sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(5), 804-812.
Kaschak, M. P., Zwaan, R. A., Aveyard, M., & Yaxley, R. H. (2006). Perception of auditory motion affects language processing. *Cognitive science*, *30*(4), 733-744.
Katz, J., & Fodor, J. (1963). The structure of a semantic theory. *Language*, *39*(2), 170-210.
Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. *Current biology*, *15*(21), 1943-1947.
Keetels, M., & Stekelenburg, J. J. (2014). Motor-induced visual motion: hand movements driving visual motion perception. *Experimental brain research*, *232*, 2865-2877.
Kelly, S., Zhou, W., Bansal, S., Peterson, M. S., & Joiner, W. M. (2019). The temporal and spatial constraints of saccade planning to double-step target displacements. *Vision research*, *163*, 1-13.

Kemmerer, D. (2015). Are the motor features of verb meanings represented in the precentral motor cortices? Yes, but within the context of a flexible, multilevel architecture for conceptual knowledge. *Psychonomic Bulletin & Review*, *22*, 1068-1075.
Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive science*, *11*(2), 201-258.
Kent, C., & Lamberts, K. (2008). The encoding–retrieval relationship: retrieval as mental simulation. *Trends in cognitive sciences*, *12*(3), 92-98.
Keogh, R., Bergmann, J., & Pearson, J. (2020). Cortical excitability controls the strength of mental imagery. *elife*, *9*, e50232.
Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805-825.
Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, *95*(2), 163-182.
Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of memory and language*, *29*(2), 133-159.
Klein, W. (2009). How time is encoded. In Klein, W., & Li, P. (eds.). *The Expression of Time.* Berlin: De Gruyter, 39-82.
Knudson, D. V., & Lindsey, C. (2014). Type I and Type II errors in correlations of various sample sizes. *Comprehensive Psychology*, *3*, 1.
Koć-Januchta, M., Höffler, T., Thoma, G. B., Prechtl, H., & Leutner, D. (2017). Visualizers versus verbalizers: Effects of cognitive style on learning with texts and pictures–An eye-tracking study. *Computers in human behavior*, *68*, 170-179.
Konen, C. S., & Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nature neuroscience*, *11*(2), 224-231.
Körner, A., Topolinski, S., & Strack, F. (2015). Routes to embodiment. *Frontiers in psychology*, *6*, 940.
Kosslyn, S. M. (1980). *Image and mind*. Cambridge MA: Harvard University Press.
Kosslyn, S. M. (1994). *Image and Brain: The resolution of the imagery debate*. Cambridge MA: MIT Press.
Kosslyn, S. M. (2005). Mental images and the Brain. *Cognitive Neuropsychology*, *22*(3-4), 333-347.
Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: evidence from studies of image scanning. *Journal of experimental psychology: Human perception and performance*, *4*(1), 47-60.
Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience*, *2*(9), 635-642.
Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford: University Press.
Kosslyn, S. M., Thompson, W. L., Sukel, K. E., & Alpert, N. M. (2005). Two types of image generation: Evidence from PET. *Cognitive, affective, & behavioral neuroscience*, *5*(1), 41-53.
Kowler, E. (2009). Attention and eye movements. In Squire, L. R. (ed.). *Encyclopedia of neuroscience*. New York: Academic Press, 605-616.
Kozhevnikov, M. (2007). Cognitive styles in the context of modern psychology: toward an integrated framework of cognitive style. *Psychological bulletin*, *133*(3), 464-481.
Kozhevnikov, M., Blazhenkova, O., & Becker, M. (2010). Trade-off in object versus spatial visualization abilities: Restriction in the development of visual-processing resources. *Psychonomic bulletin & review*, *17*(1), 29-35.
Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3-4), 243-276.
Kuperberg, G., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, *31*(1), 32-59.
Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in cognitive sciences*, *12*(2), 72-79.
Kutas, M., Federmeier, K. D., Staab, J., & Kluender, R. (2007). Language. In Cacioppo, J., Tassinary, L., & Berntson, G. (eds.). *Handbook of psychophysiology* (3rd ed.). Cambridge: University Press, 555-580.
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203-205.
Kvamme, T. L., Sandberg, K., & Silvanto, J. (2024). Mental Imagery as part of an 'Inwardly Focused' Cognitive Style. *Neuropsychologia*, 108988.
Lacey, S., & Lawson, R. (eds.) (2013). *Multisensory Imagery*. New York: Springer.
Laeng, B., Bloem, I. M., D'Ascenzo, S., & Tommasi, L. (2014). Scrutinizing visual images: The role of gaze in mental imagery and memory. *Cognition*, *131*(2), 263-283.
Laeng, B., & Teodorescu, D. S. (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive science*, *26*(2), 207-231.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive science*, *4*(2), 195-208.
Lakoff, G. (1987). *Women, Fire, and Dangerous Things. What Categories reveal about the Mind.* Chicago: University of Chicago Press.
Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature reviews neuroscience, 18*(1), 42-55.
Land, M. (2019). Eye movements in man and other animals. *Vision research*, *162*, 1-7.
Landau, B. (2017). Update on "what" and "where" in spatial language: A new division of labor for spatial terms. *Cognitive science*, *41*, 321-350.
Langacker, R. W. (1986). An introduction to cognitive grammar. *Cognitive science*, *10*(1), 1-40.
Lavie, N. (2005). Distracted and confused? Selective attention under load. *Trends in cognitive sciences*, *9*(2), 75-82.
Lemaitre, G., Houix, O., Misdariis, N., & Susini, P. (2010). Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: Applied, 16*(1), 16-32.
Lencer, R., Meyhöfer, I., Triebsch, J., Rolfes, K., Lappe, M., & Watson, T. (2021). Saccadic suppression in schizophrenia. *Scientific Reports*, *11*(1), 13133.
Levelt, W. (1989). *Speaking: from intention to articulation.* Cambridge MA: MIT Press.
Levelt, W. (1999). Producing spoken language: A blueprint of the speaker. In Brown, C., & Hagoort, P. (eds.). *The neurocognition of language*. Oxford: University Press, 83-122.
Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences*, *22*(1), 1-38.
Levinson, S. C. (1997). From outer to inner space: linguistic categories and non-linguistic thinking. *Language and Conceptualization*, *1*, 13-45.
Liman, T. G., & Zangemeister, W. H. (2012). Scanpath Eye Movements During Visual Mental Imagery in a Simulated Hemianopia Paradigm. *Journal of Eye Movement Research*, *5*(1), 1-11.
Lindsay, S., Scheepers, C., & Kamide, Y. (2013). To dash or to dawdle: Verb-associated speed of motion influences eye movements during spoken sentence comprehension. *PloS one*, *8*(6), e67187.
Liu, X. (2009). *System-level attention links cognition, perception and action: Evidence from language comprehension and eye movements.* PhD thesis (University of York).
Logie, R. H. (2003). Spatial and visual working memory: A mental workspace. In Federmeier, K. (ed.). *Psychology of learning and motivation*, *Vol. 42*. New York: Academic Press, 37-78.
Long, J. (2022). *jtools: Analysis and Presentation of Social Scientific Data*. R package version 2.2.0.
Louwerse, M., & Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. In de Vega, M., Glenberg, A., & Graesser, A. (eds.). *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford: University Press, 309-326.
Lovell, G., & Collins, D. J. (2002). Electroencephalographic differences between high and low mental imagery ability groups when learning a novel motor skill. *Journal of Human Movement Studies*, *43*, 269-296.
Lüdecke, D. (2024). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.17
Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D., (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software, 6*(60), 3139.
Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in psychology*, *3*, 54.
Lupyan, G., & Bergen, B. (2016). How language programs the mind. *Topics in cognitive science*, *8*(2), 408-424.
Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, *141*(1), 170-186.
Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of physiology-Paris*, *102*(1-3), 59-70.
Mar, R. A., & Oatley, K. (2008). The Function of Fiction Is the Abstraction and Simulation of Social Experience. *Perspectives on Psychological Science*, *3*(3), 173-192.
Marcell, M. M., Borella, D., Greene, M., Kerr, E., & Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of clinical and experimental neuropsychology*, *22*(6), 830-864.
Marconi, M., Blanco, N. D. C., Zimmer, C., & Guyon, A. (2023). Eye movements in response to different cognitive activities measured by eyetracking: a prospective study on some of the neurolinguistics programming theories. *Journal of Eye Movement Research*, *16*(2), 2.
Marks, D. F. (1973). Visual imagery differences in the recall of pictures. *British journal of psychology*, *64*(1), 17-24.
Marks, D. F. (1995). New directions for mental imagery research. *Journal of Mental Imagery, 19*(3-4), 153–167.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge MA: MIT Press.
Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, *10*(1), 29-63.
Martarelli, C. S., & Mast, F. W. (2013). Eye movements during long-term pictorial recall. *Psychological research*, *77*, 303-309.
Matlock, T. (2004). Fictive motion as cognitive simulation. *Memory & cognition*, *32*, 1389-1400.
McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, *18*(1), 1-86.
McDermott, J. (2009). The cocktail party problem. *Current Biology*, *19*(22), R1024-R1027.
McDermott, J. (2013). Audition. In Ochsner, K., & Kosslyn, S. (eds). *Oxford Handbook of Cognitive Neuroscience, Vol. 1: Core Topics.* Oxford: University Press, 135-170.
McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nature Neuroscience*, *16*(4), 493-498.
McKelvie, S. J. (1995). The VVIQ as a psychometric test of individual differences in visual imagery vividness: A critical quantitative review and plea for direction. *Journal of Mental Imagery, 19*(3-4), 1-106.
McKelvie, S. J., & Demers, E. G. (1979). Individual differences in reported visual imagery and memory performance. *British journal of psychology*, *70*(1), 51-57.
McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of learning and motivation*, *51*, 297-384.
Mechelli, A., Price, C. J., Friston, K. J., & Ishai, A. (2004). Where bottom-up meets top-down: neuronal interactions during perception and imagery. *Cerebral cortex*, *14*(11), 1256-1265.
Menenti, L., Gierhan, S. M., Segaert, K., & Hagoort, P. (2011). Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychological science*, *22*(9), 1173-1182.
Meteyard, L., Bahrami, B., & Vigliocco, G. (2007). Motion detection and motion verbs: Language affects low-level visual perception. *Psychological Science*, *18*(11), 1007-1013.
Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of memory and language*, *112*, 104092.
Micheyl, C., Carlyon, R. P., Gutschalk, A., et al. (2007). The role of auditory cortex in the formation of auditory streams. *Hearing research*, *229*(1-2), 116-131.
Micic, D., Ehrlichman, H., & Chen, R. (2010). Why do we move our eyes while trying to remember? The relationship between non-visual gaze patterns and memory. *Brain and Cognition*, *74*(3), 210-224.
Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202.
Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action.* Oxford: University Press.
Minsky, M. (1974). *A framework for representing knowledge*. Cambridge MA: MIT Press.
Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, *6*, 414-417.
Morris, L. S., Grehl, M. M., Rutter, S. B., Mehta, M., & Westwater, M. L. (2022). On what motivates us: a detailed review of intrinsic v. extrinsic motivation. *Psychological medicine*, *52*(10), 1801-1816.
Morrow, D. G. (1985). Prominent characters and events organize narrative understanding. *Journal of memory and language*, *24*(3), 304-319.
Mota-Rolim, S. A. (2020). On moving the eyes to flag lucid dreaming. *Frontiers in Neuroscience*, *14*, 361.
Moulton, S. & Kosslyn, S. (2009). Imagining predictions: mental imagery as mental emulation. *Philosophical Transactions of the Royal Society B, 364*, 1273-1280.
Munoz, D. P., Armstrong, I. T., Hampton, K. A., & Moore, K. D. (2003). Altered control of visual fixation and saccadic eye movements in attention-deficit hyperactivity disorder. *Journal of neurophysiology*, *90*(1), 503-514.
Nanay, B. (2018). Multimodal mental imagery. *Cortex*, *105*, 125-134.
Nanay, B. (2021). Unconscious mental imagery. *Philosophical Transactions of the Royal Society B*, *376*(1817), 20190689.
Näätanen, R., & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological Bulletin*, *125*(6), 826-859.
Neisser, U. (1967). *Cognitive Psychology*. New York: Appleton-Century-Crofts.
Neisser, U. (1976). *Cognition and reality: principles and implications of cognitive psychology*. San Francisco: Freeman.
Newen, A., De Bruin, L., & Gallagher, S. (eds.) (2018). *The Oxford Handbook of 4E Cognition*. Oxford: University Press.

Newtson, D., Engquist, G. A., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology, 35*(12), 847–862.
Noppeney, U. (2009). The sensory-motor theory of semantics: Evidence from functional imaging. *Language and Cognition*, *1*(2), 249-276.
Noppeney, U., Josephs, O., Hocking, J., Price, C. J., & Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cerebral Cortex*, *18*(3), 598-609.
Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior research methods*, *42*(1), 188-204.
O'Callaghan, C. (2012). Perception. In Frankish, K., & Ramsey, W. (eds.). *Cambridge handbook of cognitive science*. Cambridge: University Press, 73-91.
O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, *24*(5), 939-973.
Ohlendorf, A., Schaeffel, F., & Wahl, S. (2022). Positions of the horizontal and vertical centre of rotation in eyes with different refractive errors. *Ophthalmic and Physiological Optics*, *42*(2), 376-383.
Otero-Millan, J., Castro, J. L. A., Macknik, S. L., & Martinez-Conde, S. (2014). Unsupervised clustering method to detect microsaccades. *Journal of vision*, *14*(2), 18-18.
Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, and Winston.
Papafragou, A., Hulbert, J. & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition, 108*, 155-184.
Papafragou, A., Massey, C., & Gleitman, L. (2006). When English proposes what Greek presupposes: The cross-linguistic encoding of motion events. *Cognition*, *98*(3), B75-B87.
Pastukhov, A. (2022). *saccadr: Extract Saccades via an Ensemble of Methods Approach*, v. 0.1.3.
Pavlenko, A. (2012). Affective processing in bilingual speakers: Disembodied cognition?. *International Journal of Psychology*, *47*(6), 405-428.
Pearson, J., & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the national academy of sciences*, *112*(33), 10089.
Peltonen, V., Eronen, A., Parviainen, M., & Klapuri, A. (2001). Recognition of Everyday Auditory Scenes: Potentials, Latencies and Clues. In *Audio Engineering Society*, Convention Paper, Presented at the 110th Convention, 2001 May 12-15, Amsterdam.
Perky, C. W. (1910). An Experimental Study of Imagination. *The American Journal of Psychology*, *21*(3), 422-452.
van Petten, C., & Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia*, *33*(4), 485-508.
Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in cognitive sciences*, *11*(3), 105-110.
Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences*, *36*(4), 329-347.
Posner, M. I. (1973). Coordination of internal codes. In Chase, W. (ed.). *Visual information processing*. New York: Academic Press, 35-73.
Postle, B. R., Idzikowski, C., Sala, S. D., Logie, R. H., & Baddeley, A. D. (2006). The selective disruption of spatial working memory by eye movements. *Quarterly Journal of Experimental Psychology*, *59*(1), 100-120.
Pourcel, S. (2004). What makes path of motion salient? *Annual Meeting of the Berkeley Linguistics Society*, 505-516.
Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and brain sciences*, *22*(2), 253-279.
Pulvermüller, F. (2013). Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. *Brain and language*, *127*(1), 86-103.
Pulvermüller, F., Hauk, O., Nikulin, V. V., & Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *European Journal of Neuroscience*, *21*(3), 793-797.
Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, *103*(20), 7865-7870.
Purves, D. (ed). (2012). *Neuroscience* (2nd ed). Sunderland MA: Sinauer.
Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin, 80*(1), 1-24.
Pylyshyn, Z. W. (1981). The imagery debate: Analogue media versus tacit knowledge. *Psychological review, 88*(1), 16-45.
Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and brain sciences*, *25*(2), 157-182.
R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. [Computer program]

Radvansky, G. & Zacks, J. (2014). *Event Cognition*. Oxford: University Press.
Reilly, J., Shain, C., Borghesani, V. et al. (2025). What we mean when we say semantic: Toward a multidisciplinary semantic glossary. *Psychonomic Bulletin & Review, 32*, 243-280.
Repérant, J., Miceli, D., Vesselkin, N. P., & Molotchnikoff, S. (1989). The centrifugal visual system of vertebrates: a century-old search reviewed. *International Review of Cytology*, *118*, 115-171.
Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive science*, *31*(4), 613-643.
Richardson, D. C., Dale, R., & Spivey, M. J. (2008). Eye movements in language and cognition: A brief introduction. In Gonzalez-Marquez, M., Mittelberg, I., Coulson, S., Spivey, M. J. (eds.). *Methods in cognitive linguistics*. Amsterdam: John Benjamins, 323-344.
Richardson, D. C., Spivey, M. J., Barsalou, L. W., & McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. *Cognitive science*, *27*(5), 767-780.
Richmond, L. & Zacks, J. (2017). Constructing experience: Event models from perception to action. *Trends in cognitive sciences, 21*(12), 962-980.
Riding, R. J. (1997). On the nature of cognitive style. *Educational psychology*, *17*(1-2), 29-49.
Riemer, N. (2015). Internalist semantics. Meaning, conceptualization and expression. In Riemer, N. (ed.). *Routledge Handbook of Semantics.* New York: Routledge, 30-47.
Rissman, L., & Majid, A. (2019). Thematic roles: Core knowledge or linguistic construct? *Psychonomic Bulletin & Review*, *26*(6), 1850-1869.
Röder, B., & Rösler, F. (2003). Memory for environmental sounds in sighted, congenitally blind and late blind adults: evidence for cross-modal compensation. *International Journal of Psychophysiology*, *50*(1-2), 27-39.
Rossini, P. M., Barker, A. T., Berardelli, A., et al. (1994). Non-invasive electrical and magnetic stimulation of the brain, spinal cord and roots: basic principles and procedures for routine clinical application. Report of an IFCN committee. *Electroencephalography and clinical neurophysiology*, *91*(2), 79-92.
Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. Cambridge MA: MIT Press.
Ryan, J. D., & Shen, K. (2020). The eyes are a window into memory. *Current Opinion in Behavioral Sciences*, *32*, 1-6.
Sacks, O. (2010). *The mind's eye.* New York: Knopf.
Santos, A., Chaigneau, S. E., Simmons, W. K., & Barsalou, L. W. (2011). Property generation reflects word association and situated simulation. *Language and Cognition*, *3*(1), 83-119.
Saussure, F. de (1916). Nature of the linguistic sign. *Course in general linguistics*, 65-70.
Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 773-786.
Schank, R., & Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures* (1st ed.). Psychology Press.
Schriefers, H., Meyer, A. S., & Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of memory and language*, *29*(1), 86-102.
Schwarzkopf, D. S. (2024). What is the true range of mental imagery?. *Cortex*, *170*, 21-25.
Searle, J. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.
Searle, J. (1992). *The rediscovery of the mind*. Cambridge MA: MIT Press.
Shapiro, F. (1989). Efficacy of the eye movement desensitization procedure in the treatment of traumatic memories. *Journal of traumatic stress*, *2*(2), 199-223.
Shapiro, F., & Solomon, R. (2017). Eye movement desensitization and reprocessing therapy. In Gold, S. (ed.), *APA handbook of trauma psychology: Trauma practice*. American Psychological Association, 193-212.
Shapiro, K. A., Pascual-Leone, A., Mottaghy, F. M., Gangitano, M., & Caramazza, A. (2001). Grammatical distinctions in the left frontal cortex. *Journal of cognitive neuroscience*, *13*(6), 713-720.
Sheehan, P. W. (1967). A Shortened Form Of Betts' Questionnaire Upon Mental Imagery. *Journal of clinical psychology*, *23*(3), 386-389.
Shen, K., Bezgin, G., Selvam, R., McIntosh, A. R., & Ryan, J. D. (2016). An anatomical interface between memory and oculomotor systems. *Journal of Cognitive Neuroscience*, *28*(11), 1772-1783.
Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701-703.

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, *111*(43), E4687-E4696.
Sima, J. F. (2014). *A computational theory of visuo-spatial mental imagery*. PhD thesis (Bremen University).
Simmons, W. K., Hamann, S. B., Harenski, C. L., Hu, X. P., & Barsalou, L. W. (2008). fMRI evidence for word association and situated simulation in conceptual processing. *Journal of Physiology-Paris*, *102*(1-3), 106-119.
Singer, W. (2009). Distributed processing and temporal codes in neuronal networks. *Cognitive neurodynamics*, *3*, 189-196.
Siqueiros Sanchez, M., Falck-Ytter, T., Kennedy, D. P., Bölte, S., Lichtenstein, P., D'Onofrio, B. M., & Pettersson, E. (2020). Volitional eye movement control and ADHD traits: a twin study. *Journal of Child Psychology and Psychiatry*, *61*(12), 1309-1316.
Slobin, D. (1996). From "thought and language" to "thinking for speaking". In Gumperz, J. & Levinson, S. (eds.), *Rethinking Linguistic Relativity*. Cambridge: University Press, 70-96.
Slobin, D. (2000). Verbalized events: A dynamic approach to linguistic relativity and determinism. In Niemeier, S., & Dirven, R. (eds.). *Evidence for Linguistic Relativity*. Amsterdam: John Benjamins, 107-138.
Slobin, D. (2004). The many ways to search for a frog: Linguistic typology and the expression of motion events. In Strömqvist, S., & Verhoeven, L. (eds.). *Relating Events in Narrative. Typological and Contextual Perspectives (Vol. 2).* New York: Psychology Press, 219-257.
Smallwood, J. (2013). Distinguishing how from why the mind wanders: a process–occurrence framework for self-generated mental activity. *Psychological bulletin*, *139*(3), 519-535.
Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological Bulletin, 132*(6), 946–958.
Smilek, D., Carriere, J. S. A., & Cheyne, J. A. (2010). Out of Mind, Out of Sight: Eye Blinking as Indicator and Embodiment of Mind Wandering. *Psychological Science*, *21*(6), 786-789.
Solomon, K. O., & Barsalou, L. (2004). Perceptual simulation in property verification. *Memory & cognition*, *32*(2), 244-259.
Somers, D. C., Dale, A. M., Seiffert, A. E., & Tootell, R. B. (1999). Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. *Proceedings of the National Academy of Sciences*, *96*(4), 1663-1668.
Sowa, J. F. (1984). *Conceptual structures: information processing in mind and machine*. Boston: Addison-Wesley Longman.
Speer, N. K., & Zacks, J. M. (2005). Temporal changes as event boundaries: Processing and memory consequences of narrative time shifts. *Journal of memory and language*, *53*(1), 125-140.
Spivey, M. J., & Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological research*, *65*, 235-241.
Spivey, M., Tyler, M., Richardson, D.C. & Young, E. (2000). Eye movements during comprehension of spoken scene descriptions. *Proceedings of the Annual Meeting of the Cognitive Science Society, 22*, 487-492.
S. R. Research Ltd. (2022). *EyeLink® 1000 Plus User Manual, v. 1.0.19.* Ottawa.
Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological science*, *12*(2), 153-156.
Steindorf, L., & Rummel, J. (2020). Do your eyes give you away? A validation study of eye-movement measures used as indicators for mindless reading. *Behavior research methods*, *52*(1), 162-176.
Stutterheim, C. von (1999). How language specific are processes in the conceptualiser? In Klabunde, R., & Stutterheim, C. von (eds). *Representations and Processes in Language Production.* Wiesbaden: Deutscher Universitätsverlag, 153-179.
Stutterheim, C. von (2017). „Das ist nicht falsch, klingt aber irgendwie komisch": Prinzipien der Sprachverwendung als Teil unseres Sprachwissens – Studien zum Sprachvergleich und zu fortgeschrittenen Lernervarietäten. In Konopka, M., & Wöllstein, A. (eds.). *Grammatische Variation: Empirische Zugänge und theoretische Modellierung*. Berlin: De Gruyter, 47-64.
Stutterheim, C. von, Andermann, M., Carroll, M., Flecken, M. & Schmiedtová, B. (2012). How grammaticized concepts shape event conceptualization in language production: Insights from linguistic analysis, eye tracking data, and memory performance. *Linguistics, 50*(4), 833-867.
Stutterheim, C. von & Carroll, M. (2007). Durch die Grammatik fokussiert. *Zeitschrift für Literaturwissenschaft und Linguistik, 145*, 35-60.
Stutterheim, C. von, Carroll, M. & Klein, W. (2003). Two ways of construing complex temporal structures. In Lenz, F. (ed.), *Deictic Conceptualisation of Space, Time and Person*. Amsterdam: John Benjamins, 97-133.

Stutterheim, C. von, & Gerwien, J. (2023). Die Bedeutung sprachspezifischer Ereignisschemata für die Argumentstruktur. In Hartmann, J., & Wöllstein, A. (eds.). *Propositionale Argumente im Sprachvergleich. Theorie und Empirie*. Tübingen: Narr Francke Attempto, 265-294.
Stutterheim, C. von, Gerwien, J., Bouhaous, A., Carroll, M. & Lambert, M. (2020). What makes up a reportable event in a language? Motion events as an important test domain in linguistic typology. *Linguistics*, *58*(6), 1659-1700.
Stutterheim, C. vonStutterheim, C. von, & Nüse, R. (2003). Processes of conceptualization in language production: Language-specific perspectives and event construal. *Linguistics*, *41*(5), 851-882.
Sussman, E., Ritter, W., & Vaughan, H. G. (1999). An investigation of the auditory streaming effect using event-related brain potentials. *Psychophysiology*, *36*(1), 22-34.
Suied, C., Agus, T. R., Thorpe, S. J., Mesgarani, N., & Pressnitzer, D. (2014). Auditory gist: Recognition of very short sounds from timbre cues. *The Journal of the Acoustical Society of America*, *135*(3), 1380-1391.
Talmy, L. (1972). *Semantic structures in English and Atsugewi*. PhD thesis (University of California-Berkeley).
Talmy, L. (2000a). *Toward a cognitive semantics, Vol I: Concept structuring systems*. Cambridge MA: MIT Press.
Talmy, L. (2000b). *Toward a cognitive semantics, Vol. II: Typology and process in concept structuring*. Cambridge MA: 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.
Thagard, P. (2012). Cognitive architectures. In Frankish, K., & Ramsey, W. (eds.). *Cambridge handbook of cognitive science*. Cambridge: University Press, 50-70.
Thierry, G. (2016). Neurolinguistic relativity: how language flexes human perception and cognition. *Language Learning, 66*(3), 690-713.
Thomas, N. J. (1999). Are theories of imagery theories of imagination? An active perception approach to conscious mental content. *Cognitive science*, *23*(2), 207-245.
Tillas, A. (2014). How do ideas become general in their signification? *Baltic international yearbook of cognition, logic and communication*, *9*(1), 12-47.
Tillas, A., & Vosgerau, G. (2016). Perception, action and the notion of grounding. In Müler, V. (ed.). *Fundamental Issues of Artificial Intelligence*. Cham: Springer, 459-478.
Toomey, N., & Heo, M. (2019). Cognitive ability and cognitive style: finding a connection through resource use behavior. *Instructional Science*, *47*, 481-498.
Trueswell, J. C., & Papafragou, A. (2010). Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. *Journal of memory and language*, *63*(1), 64-82.
Truman, A., & Kutas, M. (2024). Flexible Conceptual Representations. *Cognitive science*, *48:* e13475.
Tulving, E. (1991). Concepts of human memory. In Squire, L. R., Weinberger, N. M., Lynch, G. & McGaugh, J. (eds.). *Memory organization and locus of change*. Oxford: University Press, 3-32.
Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review*, *80*(5), 352-373.
Turken, A. U., & Dronkers, N. F. (2011). The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. *Frontiers in System Neuroscience*, *5*, 1.
Tversky, B. (2005). Visuospatial Reasoning. In Holyoak, K., & Morrison, R. (eds.). *The Cambridge handbook of thinking and reasoning*. Cambridge: University Press, 209-240.
Tye, M. (1991). *The imagery debate*. Cambridge MA: MIT Press.
Tyler, L. K., & Marslen-Wilson, W. (2008). Fronto-temporal brain systems supporting spoken language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1037-1054.
Vanderveer, N. J. (1979). *Ecological acoustics: Human perception of environmental sounds*. PhD thesis (Cornell University).
Venables, W., & Ripley, B. (2002). *Modern Applied Statistics with S*. New York: Springer.
Verschooren, S., & Egner, T. (2023). When the mind's eye prevails: The Internal Dominance over External Attention (IDEA) hypothesis. *Psychonomic bulletin & review*, *30*(5), 1668-1688.
Villena-González, M., López, V., & Rodríguez, E. (2016). Orienting attention to visual or verbal/auditory imagery differentially impairs the processing of visual stimuli. *Neuroimage*, *132*, 71-78.
Visser, M., & Lambon Ralph, M. A. (2011). Differential contributions of bilateral ventral anterior temporal lobe and left anterior superior temporal gyrus to semantic processes. *Journal of cognitive neuroscience*, *23*(10), 3121-3131.
Vredeveldt, A., Hitch, G. J., & Baddeley, A. D. (2011). Eyeclosure helps memory by reducing cognitive load and enhancing visualisation. *Memory & cognition*, *39*, 1253-1263.

Vukovic, N., & Williams, J. N. (2015). Individual differences in spatial cognition influence mental simulation of language. *Cognition*, *142*, 110-122.
Wahn, B., Ferris, D. P., Hairston, W. D., & König, P. (2016). Pupil sizes scale with attentional load and task experience in a multiple object tracking task. *PloS one*, *11*(12), e0168087.
Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392-393.
Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*(3), 829-853.
Weiner, S. L., & Ehrlichman, H. (1976). Ocular motility and cognitive process. *Cognition*, *4*(1), 31-43.
Wertheimer, M. (1922/2017). *Untersuchungen zur Lehre von der Gestalt*. New York: Springer.
Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.
Wilson, G., Farrell, D., Barron, I., Hutchins, J., Whybrow, D., & Kiernan, M. D. (2018). The use of eye-movement desensitization reprocessing (EMDR) therapy in treating post-traumatic stress disorder—a systematic narrative review. *Frontiers in psychology*, *9*, 923.
Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in cognitive sciences*, *13*(12), 532-540.
Winkler, I., & Schröger, E. (2015). Auditory perceptual objects as generative models: Setting the stage for communication by sound. *Brain and language*, *148*, 1-22.
Woods, K. J., & McDermott, J. H. (2018). Schema learning for the cocktail party problem. *Proceedings of the National Academy of Sciences*, *115*(14), E3313-E3322.
Wu, D. H., Morganti, A., & Chatterjee, A. (2008). Neural substrates of processing path and manner information of a moving event. *Neuropsychologia*, *46*(2), 704-713.
Xivry, J.-J. de, Missal, M., & Lefèvre, P. (2008). A dynamic representation of target motion drives predictive smooth pursuit during target blanking. *Journal of vision*, *8*(15), 6.
Yee, E., Chrysikou, E. G., Hoffman, E., & Thompson-Schill, S. L. (2013). Manual Experience Shapes Object Representations. *Psychological Science*, *24*(6), 909-919.
Zachariou, V., Klatzky, R., & Behrmann, M. (2014). Ventral and dorsal visual stream contributions to the perception of object shape and object location. *Journal of Cognitive Neuroscience*, *26*(1), 189-209.
Zacks, J. M., & Ferstl, E. C. (2016). Discourse comprehension. In Hickok, G., & Small, S. (eds). *Neurobiology of language*. New York: Academic Press, 661-673.
Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological bulletin*, *133*(2), 273-293.
Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127*(1), 3-21.
Zeman, A., Dewar, M., & Della Sala, S. (2015). Lives without imagery–Congenital aphantasia. *Cortex*, *73*, 378-380.
Zikmund, V. (ed.) (1973). *The oculomotor system and brain functions*. London: Butterworth.
Zimmer, H. D. (2008). Visual and spatial working memory: from boxes to networks. *Neuroscience & Biobehavioral Reviews*, *32*(8), 1373-1395.
Zlatev, J., & Yangklang, P. (2004). A third way to travel: The place of Thai in motion-event typology. In Strömqvist, S., & Verhoeven, L. (eds.). *Relating Events in Narrative. Typological and Contextual Perspectives (Vol. 2).* New York: Psychology Press, 159-190.
Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic bulletin & review*, *23*, 1028-1034.
Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science*, *6*(5), 292-297.
Zwaan R. A., & Madden, C. J. (2005). Embodied Sentence Comprehension. In Pecher, D., & Zwaan R.A. (eds.). *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*. Cambridge: University Press, 224-245.
Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*(2), 162-185.

# Appendix

# Appendix A: Materials

## Appendix A1: Stimulus list

| Item number | Movement direction | Played in task phase (Exp. 1) | Verbal stimulus | English translation | Stimulus duration (verbal) | LUFS (verbal) | dBFS (verbal) | Stimulus duration (non-verbal) | LUFS (non-verbal) | dBFS (non-verbal) | Sensory question filename (Exp. 1) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Vertical | Encoding & Recall | jemand geht eine Treppe hoch | someone walks up a staircase | 1477 | -35,6 | -16 | 2868 | -33,9 | -10,6 | space_outside_4528ms.wav |
| 2 | Vertical | Encoding & Recall | es regnet stark | it's raining heavily | 1269 | -33,4 | -16 | 2728 | -37,4 | -12,6 | |
| 3 | Vertical | Encoding & Recall | ein Plastikdeckel fällt runter | a plastic lid falls down | 1633 | -40,7 | -16 | 2952 | -32,8 | -13,2 | |
| 4 | Vertical | Encoding & Recall | jemand trinkt etwas | someone drinks something | 1280 | -35,3 | -18 | 1583 | -33,4 | -9,5 | |
| 5 | Vertical | Encoding & Recall | eine Münze fällt runter | a coin falls down | 1390 | -32 | -16 | 2008 | -33 | -9,8 | |
| 6 | Vertical | Encoding & Recall | ein Wasserhahn läuft | a faucet is running | 1522 | -37,9 | -16 | 1904 | -34 | -15,6 | |
| 7 | Vertical | Encoding & Recall | ein Glas Wasser wird eingeschenkt | a glass of water is being poured | 1713 | -36,3 | -16 | 2543 | -33 | -5,1 | instrument_3950ms.wav |
| 8 | Vertical | Encoding & Recall | etwas wird in eine Glasschüssel geschüttet | something is being poured into a glass bowl | 1989 | -34,2 | -16 | 1783 | -23,1 | -12,2 | |
| 9 | Vertical | Encoding & Recall | jemand läuft eine Treppe runter | someone walks down a staircase | 1676 | -35,6 | -16 | 2325 | -33,2 | -16,5 | participant_human_2522ms.wav |
| 10 | Vertical | Encoding & Recall | ein Wasserhahn tropft | a faucet is dripping | 1469 | -35,3 | -16 | 2505 | -33 | -7 | |
| 11 | Vertical | Encoding & Recall | eine Feuerwerksrakete fliegt in die Luft | a firework rocket flies into the air | 1989 | -36,8 | -16 | 2609 | -33 | -17,1 | |
| 12 | Vertical | Encoding & Recall | Murmeln fallen auf den Boden | marbles fall onto the floor | 1615 | -36,6 | -18,8 | 2540 | -34,4 | -16,5 | |
| 13 | Vertical | Encoding & Recall | etwas fällt ins Wasser | something falls into water | 1297 | -34,1 | -16 | 1719 | -26,7 | -7 | |
| 14 | Vertical | Encoding & Recall | etwas aus Glas fällt auf den Boden | something made of glass falls onto the floor | 1833 | -35,5 | -16 | 2343 | -33 | -16,2 | |
| 15 | Vertical | Encoding & Recall | Wasser rinnt in ein Spülbecken | water runs into a sink | 1654 | -38 | -16 | 2030 | -36,7 | -7,7 | instrument_3950ms.wav |
| 16 | Vertical | Encoding & Recall | Reis wird in eine Glasschüssel geschüttet | rice is being poured into a glass bowl | 1949 | -34,4 | -16 | 2226 | -35,7 | -18,5 | |
| 17 | Horizontal | Encoding & Recall | ein Rollkoffer wird über den Boden gezogen | a rolling suitcase is being dragged across the floor | 1949 | -35,3 | -16 | 2984 | -33 | -13,9 | instrument_3950ms.wav |
| 18 | Horizontal | Encoding | jemand spielt Billard | someone is playing pool | 1281 | -33,9 | -18 | 1611 | -33 | -13,6 | |
| 19 | Horizontal | Encoding & Recall | ein Fahrrad fährt vorbei | a bicycle rides past | 1354 | -37,2 | -16 | 3913 | -39,4 | -12 | |
| 20 | Horizontal | Encoding & Recall | jemand schwimmt im Wasser | someone is swimming in the water | 1376 | -35,3 | -16 | 3448 | -34,2 | -8,3 | |
| 21 | Horizontal | Encoding | eine Murmel rollt über einen Tisch | a marble rolls across a table | 1860 | -34,6 | -16 | 2304 | -34,1 | -14,7 | |
| 22 | Horizontal | Encoding & Recall | ein Pferd trabt vorbei | a horse trots past | 1640 | -35,9 | -16 | 2918 | -35,8 | -7 | |
| 23 | Horizontal | Encoding & Recall | jemand läuft über Holzboden | someone walks across wooden flooring | 1599 | -35,6 | -16 | 2203 | -41 | -16 | |
| 24 | Horizontal | Encoding | ein Motorrad fährt vorbei | a motorcycle drives past | 1462 | -34 | -16 | 2969 | -33,1 | -17,9 | participant_animal_2522ms.wav |
| 25 | Horizontal | Encoding | jemand fegt den Boden | someone sweeps the floor | 1260 | -35,8 | -17,9 | 2750 | -35,7 | -19,9 | |
| 26 | Horizontal | Encoding | eine Autoschiebetür wird geöffnet | a car sliding door is opened | 1854 | -34,5 | -16 | 3441 | -33,4 | -7 | space_inside_4034ms.wav |
| 27 | Horizontal | Encoding | jemand fährt Skateboard | someone is skateboarding | 1266 | -33,5 | -16 | 3495 | -33 | -15,4 | |
| 28 | Horizontal | Encoding & Recall | ein Auto fährt vorbei | a car drives past | 1338 | -35,8 | -16 | 2515 | -35 | -20,6 | |
| 29 | Horizontal | Encoding & Recall | jemand spielt Tischtennis | someone is playing table tennis | 1406 | -30,8 | -15,5 | 2764 | -33,6 | -9,6 | |
| 30 | Horizontal | Encoding & Recall | jemand läuft einen Gang entlang | someone walks down a hallway | 1608 | -37,5 | -16 | 2206 | -35,5 | -11,6 | |
| 31 | Horizontal | Encoding | ein Modellflugzeug fliegt vorbei | a model airplane flies past | 1753 | -38 | -16 | 2307 | -33,6 | -17,7 | |
| 32 | Horizontal | Encoding | ein Feuerwehrauto fährt vorbei | a fire truck drives past | 1699 | -37,7 | -16 | 3306 | -32,9 | -20,5 | space_outside_4528ms.wav |

# Appendix A1: Stimulus list (cont.)

| Item number | Movement direction | Played in task phase (Exp. 1) | Verbal stimulus | English translation | Stimulus duration (verbal) | LUFS (verbal) | dBFS (verbal) | Stimulus duration (non-verbal) | LUFS (non-verbal) | dBFS (non-verbal) | Sensory question filename (Exp. 1) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 33 | No Motion | Encoding | jemand pfeift | someone whistles | 1219 | -35,6 | -16 | 1713 | -33,1 | -25,2 | |
| 34 | No Motion | Encoding & Recall | ein Baby weint | a baby is crying | 1249 | -35,4 | -16 | 2166 | -33 | -21,6 | participant_human_2522ms.wav |
| 35 | No Motion | Encoding | ein Hund bellt | a dog barks | 1210 | -36 | -16 | 1563 | -33 | -19,8 | |
| 36 | No Motion | Encoding & Recall | eine Dose wird geöffnet | a can is being opened | 1511 | -34,2 | -18 | 1937 | -33 | -14,4 | time_long_4475ms.wav |
| 37 | No Motion | Encoding & Recall | jemand hustet | someone coughs | 1169 | -34,7 | -19,4 | 1644 | -33,4 | -11,8 | time_long_4475ms.wav |
| 38 | No Motion | Encoding | eine Ente quackt | a duck quacks | 1265 | -37,4 | -16 | 2061 | -33 | -17,3 | |
| 39 | No Motion | Encoding | jemand tippt auf einer Tastatur | someone is typing on a keyboard | 1819 | -35,7 | -16 | 1507 | -34,8 | -9,3 | |
| 40 | No Motion | Encoding | ein Auto hupt | a car honks | 1189 | -40,2 | -16 | 740 | -33 | -20,8 | space_outside_4528ms.wav |
| 41 | No Motion | Encoding & Recall | jemand wählt eine Telefonnummer | someone dials a phone number | 1644 | -36 | -16 | 1850 | -33 | -20,9 | |
| 42 | No Motion | Encoding & Recall | jemand räuspert sich | someone clears their throat | 1219 | -36,4 | -16 | 1184 | -33,1 | -14 | instrument_3950ms.wav |
| 43 | No Motion | Encoding | ein Pferd wiehert | a horse neighs | 1227 | -34,1 | -16 | 2323 | -33 | -20 | |
| 44 | No Motion | Encoding | ein Herz schlägt | a heart is beating | 1274 | -33 | -16 | 2102 | -34,3 | -19,7 | time_long_4475ms.wav |
| 45 | No Motion | Encoding & Recall | eine Katze miaut | a cat meows | 1282 | -35,2 | -16 | 2087 | -33 | -22,2 | participant_human_2522ms.wav |
| 46 | No Motion | Encoding | es klingelt an der Tür | the doorbell rings | 1295 | -32,9 | -16 | 1769 | -33 | -15,3 | |
| 47 | No Motion | Encoding & Recall | eine Uhr tickt | a clock is ticking | 1260 | -31,5 | -15,4 | 2596 | -33 | -11,3 | time_long_4475ms.wav |
| 48 | No Motion | Encoding & Recall | ein Auto wird gestartet | a car is being started | 1440 | -34,2 | -16 | 1941 | -33 | -14,1 | |
| 49 | Horizontal | Recall | jemand mäht den Rasen | someone is mowing the lawn | 1206 | -35,6 | -16 | 3490 | -32,6 | -15,8 | |
| 50 | Horizontal | Recall | jemand zieht einen schweren Stein über den Boden | someone drags a heavy stone across the floor | 1989 | -34,3 | -16 | 2031 | -33 | -16,5 | |
| 51 | Horizontal | Recall | jemand schreibt etwas an eine Tafel | someone writes something on a chalkboard | 1835 | -36,6 | -16 | 1994 | -37,6 | -8,8 | |
| 52 | Horizontal | Recall | ein Hund läuft über Parkettboden | a dog walks across a parquet floor | 1939 | -35,5 | -16 | 2643 | -33 | -12,9 | |
| 53 | Horizontal | Recall | mehrere Leute marschieren | several people are marching | 1551 | -32,8 | -16 | 2781 | -33 | -12 | |
| 54 | Horizontal | Recall | ein Vorhang wird aufgezogen | a curtain is being drawn open | 1676 | -34,8 | -16 | 2999 | -33 | -21 | |
| 55 | Horizontal | Recall | jemand rennt über Schotter | someone runs over gravel | 1279 | -35,1 | -16 | 2197 | -42,3 | -16,3 | |
| 56 | Horizontal | Recall | jemand schießt einen Ball gegen eine Wand | someone kicks a ball against a wall | 1879 | -34,3 | -16 | 2337 | -34 | -12,1 | |
| 57 | No Motion | Recall | mehrere Krähen krähen | several crows are cawing | 1492 | -33,6 | -16 | 1413 | -33 | -22,6 | |
| 58 | No Motion | Recall | ein Handy vibriert | a cell phone vibrates | 1304 | -34,1 | -16 | 1149 | -28,8 | -18 | |
| 59 | No Motion | Recall | jemand isst etwas knuspriges | someone is eating something crunchy | 1730 | -34,9 | -16 | 1928 | -33 | -12,4 | |
| 60 | No Motion | Recall | ein Hund winselt | a dog whines | 1277 | -36,3 | -17,9 | 1837 | -33,4 | -19,6 | |
| 61 | No Motion | Recall | Grillen zirpen | crickets are chirping | 1231 | -32,5 | -16 | 1775 | -33 | -20,3 | |
| 62 | No Motion | Recall | ein Huhn gackert | a chicken clucks | 1218 | -38,1 | -16 | 1133 | -33 | -21,4 | |
| 63 | No Motion | Recall | Vögel zwitschern | birds are chirping | 1224 | -32,2 | -16 | 1925 | -37,9 | -19,2 | |
| 64 | No Motion | Recall | jemand klickt mit einer Computermaus | someone clicks a computer mouse | 1940 | -36,4 | -16 | 1654 | -37 | -10 | |

# Appendix A2: Source files of auditory stimuli

| Item number | Filename (verbal stimulus) | Filename (non-verbal stimulus) | Source URL (non-verbal stimulus) | Source sound creator (freesound.org username) | URL accessed on |
|---|---|---|---|---|---|
| 1 | verbal_v_steigen_Treppe_hinauf_1477ms.wav | _v_steigen_Treppe_hinauf_2868ms_freesound-210430.wav | https://freesound.org/people/qubodup/sounds/210430/ | qubodup | 8th June, 2025 |
| 2 | verbal_v_regnen_1269ms.wav | _v_regnen_2728ms_freesound-2523.wav | https://freesound.org/people/RHumphries/sounds/2523/ | RHumphries | 8th June, 2025 |
| 3 | verbal_v_fallen_Plastik_Boden_1633ms.wav | _v_fallen_Plastik_Boden_2952ms_dd.wav | – | DD[1] | – |
| 4 | verbal_v_schlucken_1280ms.wav | _v_schlucken_1583ms_dd.wav | – | DD | – |
| 5 | verbal_v_fallen_Muenze_1390ms.wav | _v_fallen_Muenze_2008ms_freesound-17502.wav | https://freesound.org/people/Jace/sounds/17502/ | Jace | 8th June, 2025 |
| 6 | verbal_v_laufen_Wasser_1522ms.wav | _v_laufen_Wasser_1904ms_freesound-451761.wav | https://freesound.org/people/florianreichelt/sounds/451761/ | florianreichelt | 8th June, 2025 |
| 7 | verbal_v_eingiessen_1713ms.wav | _v_eingiessen_2543ms_freesound-154439.wav | https://freesound.org/people/lucaslara/sounds/154439/ | lucaslara | 8th June, 2025 |
| 8 | verbal_v_fuellen_MMs_Schuessel_1989ms.wav | _v_fuellen_M&Ms_Schuessel_1783ms_dd.wav | – | DD | – |
| 9 | verbal_v_steigen_Treppe_hinab_1676ms.wav | _v_steigen_Treppe_hinab_2325ms_dd.wav | – | DD | – |
| 10 | verbal_v_tropfen_Wasserhahn_1469ms.wav | _v_tropfen_Wasserhahn_2505ms_dd.wav | – | DD | – |
| 11 | verbal_v_schiessen_Feuerwerk_1989ms.wav | _v_schiessen_Feuerwerk_2609ms_freesound-455546-335983.wav | https://freesound.org/people/Kinoton/sounds/455546/; [2] https://freesound.org/people/zihengc/sounds/335983/; | Kinoton; zihengc; | 8th June, 2025 |
| 12 | verbal_v_schuetten_Murmeln_1615ms.wav | _v_schuetten_Murmeln_2540ms_freesound-157688.wav | https://freesound.org/people/urupin/sounds/157688/ | urupin | 8th June, 2025 |
| 13 | verbal_v_fallen_Stein_Wasser_1297ms.wav | _v_fallen_Stein_Wasser_1719ms_freesound-68730.wav | https://freesound.org/people/mikaelfernstrom/sounds/68730/ | mikaelfernstrom | 8th June, 2025 |
| 14 | verbal_v_fallen_Weinglas_zerbrechen_1833ms.wav | _v_fallen_Weinglas_zerbrechen_2343ms_freesound-432649.wav | https://freesound.org/people/DigPro120/sounds/432649/ | DigPro120 | 8th June, 2025 |
| 15 | verbal_v_rinnen_Wasser_1654ms.wav | _v_rinnen_Wasser_2030ms_freesound-360735.wav | https://freesound.org/people/WasabiWielder/sounds/360735/ | WasabiWielder | 8th June, 2025 |
| 16 | verbal_v_schuetten_Reis_Schuessel_1949ms.wav | _v_schuetten_Reis_Schuessel_2226ms_dd.wav | – | DD | – |
| 17 | verbal_h_ziehen_Trolley_Boden_1949ms.wav | _h_ziehen_Trolley_Boden_2984ms_freesound-402528.wav | https://freesound.org/people/yettopetto/sounds/402528/ | yettopetto | 8th June, 2025 |
| 18 | verbal_h_schiessen_Billardkugel_1281ms.wav | _h_schiessen_Billardkugel_1611ms_freesound-366337.wav | https://freesound.org/people/et_graham/sounds/366337/ | et_graham | 8th June, 2025 |
| 19 | verbal_h_vorbeifahren_Fahrrad_1354ms.wav | _h_vorbeifahren_Fahrrad_3913ms_freesound-536219.wav | https://freesound.org/people/Ambientsoundapp/sounds/536219/ | Ambientsoundapp | 8th June, 2025 |
| 20 | verbal_h_schwimmen_1376ms.wav | _h_schwimmen_3448ms_freesound-530154.wav | https://freesound.org/people/tbsounddesigns/sounds/530154/ | tbsounddesigns | 8th June, 2025 |
| 21 | verbal_h_rollen_Murmel_1860ms.wav | _h_rollen_Murmel_2304ms_dd.wav | – | DD | – |
| 22 | verbal_h_reiten_Pferd_1640ms.wav | _h_reiten_Pferd_2918ms_freesound-166099.wav | https://freesound.org/people/Zabuhailo/sounds/166099/ | Zabuhailo | 8th June, 2025 |
| 23 | verbal_h_laufen_Gang_Mann_1599ms.wav | _h_laufen_Gang_Mann_2203ms_dd.wav | – | DD | – |

# Appendix A2: Source files of auditory stimuli (cont.)

| Item number | Filename (verbal stimulus) | Filename (non-verbal stimulus) | Source URL (non-verbal stimulus) | Source sound creator (freesound.org username) | URL accessed on |
|---|---|---|---|---|---|
| 24 | verbal_h_vorbeifahren_Motorrad_1462ms.wav | _h_vorbeifahren_Motorrad_2969ms_freesound-1046.wav | https://freesound.org/people/RHumphries/sounds/1046/ | RHumphries | 8th June, 2025 |
| 25 | verbal_h_kehren_1260ms.wav | _h_kehren_2750ms_dd.wav | – | DD | – |
| 26 | verbal_h_Schiebetuer_Auto_1854ms.wav | _h_Schiebetuer_Auto_3441ms_freesound-332853.wav | https://freesound.org/people/YleArkisto/sounds/332853/ | YleArkisto | 8th June, 2025 |
| 27 | verbal_h_fahren_Skateboard_1266ms.wav | _h_fahren_Skateboard_3495ms_freesound-348329.wav | https://freesound.org/people/BlackNeon1234/sounds/348329/ | BlackNeon1234 | 8th June, 2025 |
| 28 | verbal_h_vorbeifahren_Auto_1338ms.wav | _h_vorbeifahren_Auto_2515ms_freesound-377003.wav | https://freesound.org/people/atks_/sounds/377003/ | atks_ | 8th June, 2025 |
| 29 | verbal_h_spielen_Tischtennis_1406ms.wav | _h_spielen_Tischtennis_2764ms_freesound-64921.wav | https://freesound.org/people/juskiddink/sounds/64921/ | juskiddink | 8th June, 2025 |
| 30 | verbal_h_laufen_Gang_Frau_1608ms.wav | _h_laufen_Gang_Frau_2206ms_freesound-368834.wav | https://freesound.org/people/georgisound/sounds/368834/ | georgisound | 8th June, 2025 |
| 31 | verbal_h_vorbeifliegen_Flugzeug_1753ms.wav | _h_vorbeifliegen_Flugzeug_2307ms_freesound-401157.wav | https://freesound.org/people/granizo/sounds/401157/ | granizo | 8th June, 2025 |
| 32 | verbal_h_vorbeifahren_Krankenwagen_1699ms.wav | _h_vorbeifahren_Krankenwagen_3306ms_freesound-541595.wav | https://freesound.org/people/Breviceps/sounds/541595/ | Breviceps | 8th June, 2025 |
| 33 | verbal_f_pfeifen_1219ms.wav | _f_pfeifen_1713ms_dd.wav | – | DD | – |
| 34 | verbal_f_weinen_1249ms.wav | _f_weinen_2166ms_freesound-398552.wav | https://freesound.org/people/gumballworld/sounds/398552/ | gumballworld | 8th June, 2025 |
| 35 | verbal_f_bellen_1210ms.wav | _f_bellen_1563ms_freesound-199261.wav | https://freesound.org/people/felix.blume/sounds/199261/ | felix.blume | 8th June, 2025 |
| 36 | verbal_f_oeffnen_Dose_1511ms.wav | _f_oeffnen_Dose_1937ms_freesound-175025.wav | https://freesound.org/people/MakoFox/sounds/175025/ | MakoFox | 8th June, 2025 |
| 37 | verbal_f_husten_1169ms.wav | _f_husten_1644ms_dd.wav | – | DD | – |
| 38 | verbal_f_quacken_Ente_1265ms.wav | _f_quacken_Ente_2061ms_freesound-185134.wav | https://freesound.org/people/dobroide/sounds/185134/ | dobroide | 8th June, 2025 |
| 39 | verbal_f_tippen_1819ms.wav | _f_tippen_1507ms_dd.wav | – | DD | – |
| 40 | verbal_f_hupen_1189ms.wav | _f_hupen_740ms_freesound-434878.wav | https://freesound.org/people/MicktheMicGuy/sounds/434878/ | MicktheMicGuy | 8th June, 2025 |
| 41 | verbal_f_anrufen_1644ms.wav | _f_anrufen_1850ms_dd.wav | – | DD | – |
| 42 | verbal_f_raeuspern_1219ms.wav | _f_raeuspern_1184ms_dd.wav | – | DD | – |
| 43 | verbal_f_wiehern_1227ms.wav | _f_wiehern_2323ms_freesound-59569.wav | https://freesound.org/people/3bagbrew/sounds/59569/ | 3bagbrew | 8th June, 2025 |
| 44 | verbal_f_pochen_Herz_1274ms.wav | _f_pochen_Herz_2102ms_freesound-535478.wav | https://freesound.org/people/Moulaythami/sounds/535478/ | Moulaythami | 8th June, 2025 |
| 45 | verbal_f_miauen_1282ms.wav | _f_miauen_2087ms_freesound-365076.wav | https://freesound.org/people/justiiiiin/sounds/365076/ | justiiiiin | 8th June, 2025 |
| 46 | verbal_f_klingeln_1295ms.wav | _f_klingeln_1769ms_freesound-275072.wav | https://freesound.org/people/kwahmah_02/sounds/275072/ | kwahmah_02 | 8th June, 2025 |
| 47 | verbal_f_ticken_1260ms.wav | _f_ticken_2596ms_freesound-405423-237210.wav | https://freesound.org/people/straget/sounds/405423/; https://freesound.org/people/Mortifreshman/sounds/237210/; | straget; Mortifreshman; | 8th June, 2025 |

| Item number | Filename (verbal stimulus) | Filename (non-verbal stimulus) | Source URL (non-verbal stimulus) | Source sound creator (freesound.org username) | URL accessed on |
|---|---|---|---|---|---|
| 48 | verbal_f_starten_1440ms.wav | _f_starten_1941ms_freesound-326052.wav | https://freesound.org/people/botha9johann/sounds/326052/ | botha9johann | 8th June, 2025 |
| 49 | verbal_h_maehen_Rasen_1206ms.wav | _h_maehen_Rasen_3490ms_freesound-318682.wav | https://freesound.org/people/petebuchwald/sounds/318682/ | petebuchwald | 8th June, 2025 |
| 50 | verbal_h_ziehen_Stein_Boden_1989ms.wav | _h_ziehen_Stein_Boden_2031ms_freesound-243699.wav | https://freesound.org/people/ertfelda/sounds/243699/ | ertfelda | 8th June, 2025 |
| 51 | verbal_h_schreiben_Tafel_1835ms.wav | _h_schreiben_Tafel_1994ms_freesound-332287.wav | https://freesound.org/people/Sirderf/sounds/332287/ | Sirderf | 8th June, 2025 |
| 52 | verbal_h_laufen_Hund_1939ms.wav | _h_laufen_Hund_2643ms_freesound-518166.wav | https://freesound.org/people/Elenalostale/sounds/518166/ | Elenalostale | 8th June, 2025 |
| 53 | verbal_h_marschieren_Soldaten_1551ms.wav | _h_marschieren_Soldaten_2781ms_freesound-200321.wav | https://freesound.org/people/WebbFilmsUK/sounds/200321/ | WebbFilmsUK | 8th June, 2025 |
| 54 | verbal_h_ziehen_Vorhang_1676ms.wav | _h_ziehen_Vorhang_2999ms_freesound-51138.wav | https://freesound.org/people/RutgerMuller/sounds/51138/ | RutgerMuller | 8th June, 2025 |
| 55 | verbal_h_rennen_Schotter_1279ms.wav | _h_rennen_Schotter_2197ms_freesound-456038.wav | https://freesound.org/people/florianreichelt/sounds/456038/ | florianreichelt | 8th June, 2025 |
| 56 | verbal_h_schiessen_Fussball_Wand_1879ms.wav | _h_schiessen_Fussball_Wand_2337ms_freesound-190796-543167.wav | https://freesound.org/people/jiserte/sounds/190796/; https://freesound.org/people/1skyland/sounds/543167/; | jiserte; 1skyland; | 8th June, 2025 |
| 57 | verbal_f_kraehen_1492ms.wav | _f_kraehen_1413ms_freesound-353432.wav | https://freesound.org/people/Dave_Girtsman/sounds/353432/ | Dave_Girtsman | 8th June, 2025 |
| 58 | verbal_f_vibrieren_1304ms.wav | _f_vibrieren_1149ms_freesound-179012.wav | https://freesound.org/people/SmartWentCody/sounds/179012/ | SmartWentCody | 8th June, 2025 |
| 59 | verbal_f_spitzen_1730ms.wav | _f_spitzen_1928ms_dd.wav | – | DD | – |
| 60 | verbal_f_winseln_1277ms.wav | _f_winseln_1837ms_freesound-160475.wav | https://freesound.org/people/unfa/sounds/160475/ | unfa | 8th June, 2025 |
| 61 | verbal_f_zirpen_1231ms.wav | _f_zirpen_1775ms_freesound-53380.wav | https://freesound.org/people/eric5335/sounds/53380/ | eric5335 | 8th June, 2025 |
| 62 | verbal_f_gackern_1218ms.wav | _f_gackern_1133ms_freesound-316921.wav | https://freesound.org/people/Rudmer_Rotteveel/sounds/316921/ | Rudmer_Rotte... | 8th June, 2025 |
| 63 | verbal_f_zwitschern_1224ms.wav | _f_zwitschern_1925ms_dd.wav | – | DD | – |
| 64 | verbal_f_klicken_Maus_1940.wav | _f_klicken_Maus_1654ms_freesound-355324.wav | https://freesound.org/people/NWSP/sounds/355324/ | NWSP | 8th June, 2025 |

[1]: These sounds were recorded by the author (Danny Dirker).
[2]: Two source sounds were mixed for stimulus design.

# Appendix A3: Sensory questions

| Category | Question | English translation | Duration | Filename |
|---|---|---|---|---|
| Temporal extension | Erstreckt sich das Ereignis für gewöhnlich über einen relativ langen Zeitraum? | Does this event usually span a relatively long period of time? | 4475 | time_long_4475ms.wav |
| Typical location | Würde man dieses Ereignis typischerweise außerhalb eines Gebäudes beobachten? | Would one typically observe this event outside a building? | 4528 | space_outside_4528ms.wav |
| | Würde man dieses Ereignis typischerweise in einem Gebäude beobachten? | Would one typically observe this event in a building? | 4034 | space_inside_4034ms.wav |
| Typical participant | Ist an diesem Ereignis ein Mensch beteiligt? | Is a human being involved in this event? | 2522 | participant_human_2522ms.wav |
| | Ist an diesem Ereignis ein Tier beteiligt? | Is an animal involved in this event? | | participant_animal_2522ms.wav |
| Instrument use | Handelt es sich um ein Ereignis, bei dem eine Person ein Hilfsmittel benutzt? | Is this an event in which a person uses some sort of tool? | 3950 | instrument_3950ms.wav |

# Appendix A4: Verbatim task instructions

| | Task instructions printed on screen | English translation |
|---|---|---|
| **Experiment 1 – non-verbal condition** | Spielablauf: Geräusch-Memory<br><br>Das Spiel beginnt mit einer Trainingsrunde, die Sie mit dem Spielablauf vertraut macht. Danach spielen Sie 4 Runden Geräusch-Memory. Jede Runde besteht aus zwei Phasen:<br><br>In der ersten Phase hören Sie 12 unterschiedliche Geräusche nacheinander. Ihre Aufgabe ist, sich diese Geräusche zu merken. Alle Geräusche lassen auf ein Ereignis oder eine Handlung schließen. Stellen Sie sich die Ereignisse bzw. Handlungen vor. Ab und zu werden Ihnen Fragen zu den Ereignissen gestellt. Sie beantworten die Frage mit Ja, indem Sie den grünen Punkt mit Ihrem Blick fixieren, und mit Nein, indem Sie den roten Punkt mit Ihrem Blick fixieren. Antworten Sie intuitiv und denken Sie nicht zu lange nach.<br><br>In der zweiten Phase hören Sie wieder 12 unterschiedliche Geräusche. Die meisten davon haben Sie in der ersten Phase schon gehört und sich gemerkt, aber manche sind auch neu. Ihre Aufgabe ist nun, bei neuen Geräuschen einen Knopf auf dem Controller zu drücken. Drücken Sie also bei denjenigen Geräuschen, die Sie in der ersten Phase noch nicht gehört haben. Dabei ist es egal, welchen Knopf Sie drücken. Wenn Sie ein Geräusch hören, das in der ersten Phase vorgekommen ist, tun Sie einfach nichts und das nächste wird wenigen Sekunden später automatisch abgespielt. Fragen werden Ihnen hier nicht gestellt.<br><br>Anmerkungen<br>• Jedes Geräusch wird nur einmal abgespielt und kann nicht wiederholt werden. Die Reihenfolge der Geräusche ist zufällig und spielt keine Rolle. Zwischen den Geräuschen bestehen keine Zusammenhänge.<br>• Sie müssen sich die Geräusche immer nur innerhalb einer Runde merken. Das heißt, dass Sie sich nicht in Runde 4 an ein Geräusch aus Runde 1 erinnern sollen.<br>• Wenn Sie ein Geräusch nicht erkennen, versuchen Sie bitte trotzdem, es sich zu merken.<br>• Die Fragen beziehen sich immer nur auf das direkt davor abgespielte Geräusch, nicht auf mehrere.<br>• Bleiben Sie mit Ihrem Blick bitte innerhalb des grauen Quadrats und vermeiden Sie übermäßig häufiges Blinzeln. | Game Procedure: Sound Memory<br><br>The game begins with a training round to familiarize you with the procedure. After that, you will play 4 rounds of Sound Memory. Each round consists of two phases:<br><br>In the first phase, you will hear 12 different sounds one after the other. Your task is to remember these sounds. Each sound suggests an event or an action. Imagine the corresponding events or actions. From time to time, you will be asked questions about the events. Answer “yes” by fixating on the green dot with your gaze, and “no” by fixating on the red dot with your gaze. Answer intuitively and don’t overthink your response.<br><br>In the second phase, you will again hear 12 different sounds. Most of these were already presented in the first phase and should be remembered, but some are new. Your task now is to press a button on the controller when you hear a new sound. In other words, press a button for any sound that was not played in the first phase. It doesn’t matter which button you press. If you hear a sound that was presented in the first phase, simply do nothing — the next sound will play automatically after a few seconds. No questions will be asked during this phase.<br><br>Note<br>• Each sound is played only once and cannot be repeated. The order of the sounds is random and irrelevant. There are no connections between the sounds.<br>• You only need to remember the sounds within a single round. That means you are not expected to recall a sound from Round 1 during Round 4.<br>• If you do not recognize a sound, please try to remember it anyway.<br>• The questions always refer only to the sound played directly before, not to multiple sounds.<br>• Please keep your gaze within the gray square and avoid excessive blinking. |

| | | |
|---|---|---|
| **Experiment 1 – verbal condition** | Spielablauf: Ereignis-Memory<br><br>Das Spiel beginnt mit einer Trainingsrunde, die Sie mit dem Spielablauf vertraut macht. Danach spielen Sie 4 Runden Ereignis-Memory. Jede Runde besteht aus zwei Phasen:<br><br>In der ersten Phase hören Sie 12 unterschiedliche Ereignisbeschreibungen nacheinander. Ihre Aufgabe ist, sich diese Ereignisse oder Handlungen vorzustellen und zu merken. Denken Sie z. B. daran, wie sich das Ereignis anhört. Ab und zu werden Ihnen Verständnisfragen gestellt. Diese beantworten Sie mit Ja, indem Sie den grünen Punkt mit Ihrem Blick fixieren, und mit Nein, indem Sie den roten Punkt mit Ihrem Blick fixieren. Antworten Sie intuitiv und denken Sie nicht zu lange nach.<br><br>In der zweiten Phase hören Sie wieder 12 unterschiedliche Ereignisbeschreibungen. Die meisten davon haben Sie in der ersten Phase schon gehört und sich gemerkt, aber manche sind auch neu. Ihre Aufgabe ist nun, bei neuen Ereignissen einen Knopf auf dem Controller zu drücken. Drücken Sie also bei denjenigen Ereignissen, die Sie in der ersten Phase noch nicht gehört haben. Dabei ist es egal, welchen Knopf Sie drücken. Wenn Sie ein Ereignis hören, das in der ersten Phase vorgekommen ist, tun Sie einfach nichts und das nächste wird in wenigen Sekunden automatisch abgespielt. Fragen werden Ihnen hier nicht gestellt.<br><br>Anmerkungen<br>• Jede Ereignisbeschreibung wird nur einmal abgespielt und kann nicht wiederholt werden. Die Abspielreihenfolge ist zufällig und spielt keine Rolle, weil zwischen den Ereignissen keine Zusammenhänge bestehen.<br>• Die Fragen beziehen sich immer nur auf das direkt davor abgespielte Ereignis, nicht auf mehrere.<br>• Sie müssen sich die Ereignisse immer nur innerhalb einer Runde merken. Das heißt, dass Sie sich nicht in Runde 4 an ein Ereignis aus Runde 1 erinnern sollen.<br>• Manche Ereignisse beginnen ähnlich – hören Sie also gut hin und reagieren sie nicht vorschnell.<br>• Bleiben Sie mit Ihrem Blick bitte innerhalb des grauen Quadrats und vermeiden Sie es, übermäßig häufig zu blinzeln. | Game Procedure: Event Memory<br><br>The game begins with a training round to familiarize you with the procedure. After that, you will play 4 rounds of Event Memory. Each round consists of two phases:<br><br>In the first phase, you will hear 12 different event descriptions one after the other. Your task is to imagine and remember these events or actions. For example, think about what the event might sound like. From time to time, you will be asked comprehension questions. Answer “yes” by fixating on the green dot with your gaze, and “no” by fixating on the red dot with your gaze. Answer intuitively and don’t overthink your response.<br><br>In the second phase, you will again hear 12 different event descriptions. Most of these were already presented in the first phase and should be remembered, but some are new. Your task now is to press a button on the controller when you hear a new event. In other words, press a button for any event that was not described in the first phase. It doesn’t matter which button you press. If you hear an event that was presented in the first phase, simply do nothing — the next one will play automatically after a few seconds. No questions will be asked during this phase.<br><br>Note<br>• Each event description is played only once and cannot be repeated. The order of presentation is random and does not matter, as there are no connections between the events.<br>• The questions always refer only to the event that was played directly before, not to multiple events.<br>• You only need to remember the events within a single round. That means you are not expected to recall an event from Round 1 during Round 4.<br>• Some events may begin similarly — so listen carefully and don’t respond too quickly.<br>• Please keep your gaze within the gray square and avoid excessive blinking. |
| **Experiment 2** | Beschreibungsaufgabe<br><br>In dieser letzten Aufgabe sollen Sie beschreiben, was Sie hören. Über 4 Runden werden Ihnen nun noch mal die Geräusche vorgespielt, die Sie im Memory-Spiel gehört haben. Sie sollen sagen, welches Ereignis oder welche Handlung sich in den jeweiligen Geräuschen vollzieht.<br><br>Ablauf<br>Sie hören ein Geräusch. Sobald Sie es beschreiben können, drücken Sie einen Knopf auf dem Controller, um die Sprachaufnahme zu starten. Sprechen Sie dann klar und deutlich, aber spontan und natürlich in das Mikrofon. Die Sprachaufnahme endet automatisch und das nächste Geräusch wird Ihnen vorgespielt.<br><br>Drücken Sie [die mittlere Taste auf dem Controller], um sich Beispiele anzuhören.<br><br>Anmerkungen<br>• Beschreiben Sie in Ihrer Antwort einfach, was passiert. Sie müssen keinen kompletten Satz sprechen, aber nutzen Sie genügend Wörter, um konkret zu sein. Stellen Sie sich vor, Sie würden telefonieren und müssten beschreiben, was Sie gerade hören.<br>• Wenn Sie das Geräusch nicht erkannt haben, sagen Sie »Keine Ahnung«.<br>• Sie müssen für die Sprachaufnahme nicht unbedingt warten - unterbrechen Sie die Geräusche einfach mit einem Knopfdruck, sobald Sie sprechen können.<br>• Denken Sie daran, immer erst zu drücken, um die Sprachaufnahme zu starten. Sonst nimmt das Programm Ihre Stimme nicht auf. Sie müssen den Knopf nicht gedrückt halten, während Sie sprechen.<br>• Bleiben Sie mit Ihrem Blick bitte innerhalb des grauen Quadrats und vermeiden Sie übermäßig häufiges Blinzeln. | Description Task<br><br>In this final task, you will describe what you hear. Over the course of 4 rounds, the sounds from the Memory Game will be played again. Your task is to say what event or action is taking place in each sound.<br><br>Procedure<br>You will hear a sound. As soon as you are able to describe it, press a button on the controller to start the voice recording. Then speak clearly and distinctly, but spontaneously and naturally into the microphone. The recording will stop automatically, and the next sound will be played.<br><br>Press [the middle button on the controller] to listen to examples.<br><br>Note<br>• Simply describe what is happening in the sound. You don’t need to speak in full sentences, but please use enough words to be specific. Imagine you are on the phone and need to describe what you're hearing.<br><br>• If you don’t recognize the sound, say “No idea.”<br><br>• You don’t have to wait for the sound to finish—interrupt it with a button press as soon as you’re ready to speak.<br><br>• Remember to press the button *before* you start speaking to begin the recording. Otherwise, your voice won’t be recorded. You don’t need to hold the button down while speaking.<br>• Please keep your gaze within the gray square and avoid excessive blinking. |

## Appendix B: Data tables and analysis results

### Appendix B1: Distribution of analyzed trials across conditions (Exp. 1)

| | | | Non-verbal | | Verbal | |
|---|---|---|---|---|---|---|
| | | | Encoding | Recall | Encoding | Recall |
| Travel distance | Movement direction | Vertical | 602 | 569 | 619 | 596 |
| | | Horizontal | 612 | 316 | 614 | 309 |
| | | No motion | 595 | 317 | 597 | 317 |
| | Total | | 1809 | 1202 | 1830 | 1222 |
| | | | 3011 | | 3052 | |
| | | | 6063 | | | |

| | | | Non-verbal | | Verbal | |
|---|---|---|---|---|---|---|
| | | | Encoding | Recall | Encoding | Recall |
| Saccade rate | Movement direction | Vertical | 561 | 546 | 559 | 569 |
| | | Horizontal | 560 | 298 | 564 | 293 |
| | | No motion | 570 | 304 | 598 | 314 |
| | Total | | 1691 | 1148 | 1721 | 1176 |
| | | | 2839 | | 2897 | |
| | | | 5736 | | | |

## Appendix B2: Full model summary (Experiment 1)

| | Travel distance(x) | | | Travel distance(y) | | | Saccade rate | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimate | CI | p | Estimate | CI | p | Estimate | CI | p |
| (Intercept) | 1.40 | 1.17 – 1.62 | <0.001 | 0.28 | 0.08 – 0.48 | 0.007 | -0.09 | -0.22 – 0.03 | 0.14 |
| Movement direction [vertical] | 0.16 | 0.12 – 0.20 | <0.001 | 0.12 | 0.08 – 0.17 | <0.001 | 0.02 | -0.01 – 0.05 | 0.23 |
| Movement direction [horizontal] | 0.28 | 0.24 – 0.33 | <0.001 | 0.19 | 0.14 – 0.24 | <0.001 | 0.01 | -0.02 – 0.04 | 0.47 |
| Visualization intensity [5] | -0.37 | -0.75 – 0.00 | 0.05 | -0.10 | -0.44 – 0.24 | 0.559 | -0.20 | -0.41 – 0.01 | 0.06 |
| Visualization intensity [3] | 0.27 | -0.16 – 0.71 | 0.22 | 0.16 | -0.24 – 0.55 | 0.435 | -0.12 | -0.36 – 0.12 | 0.32 |
| Visualization intensity [2] | 0.49 | -0.06 – 1.04 | 0.08 | 0.34 | -0.15 – 0.84 | 0.172 | 0.12 | -0.18 – 0.42 | 0.44 |
| Trial duration (z-scored) | 0.13 | 0.11 – 0.15 | <0.001 | 0.10 | 0.08 – 0.12 | <0.001 | 0.01 | -0.00 – 0.02 | 0.06 |
| Task phase [recall] | -0.21 | -0.25 – -0.17 | <0.001 | -0.09 | -0.14 – -0.05 | <0.001 | 0.32 | 0.30 – 0.35 | <0.001 |
| Stimulus modality [verbal] | -0.71 | -0.74 – -0.67 | <0.001 | -0.50 | -0.54 – -0.46 | <0.001 | 0.13 | 0.11 – 0.16 | <0.001 |
| Trial number | – | – | – | – | – | – | 0.002 | 0.001–0.002 | <0.001 |
| **Random effects and model fit** | | | | | | | | | |
| Resid. variance (σ2) | 0.46 | | | 0.56 | | | 0.21 | | |
| Subject-level intercept | 1.03 | | | 0.86 | | | 0.08 | | |
| Subject-level slope for rate | 2.47 | | | 1.45 | | | – | | |
| Correlation subject-rate | -0.87 | | | -0.87 | | | – | | |
| ICC | 0.69 | | | 0.61 | | | 0.26 | | |
| R2 (marg./cond.) | 0.14 / 0.73 | | | 0.07 / 0.63 | | | 0.14 / 0.36 | | |

## Appendix B2: Full model summary (Experiment 1)

### Wald-test results

| | Travel distance(x) | | Travel distance(y) | | Saccade rate | |
|---|---|---|---|---|---|---|
| | Estimate (*df*) | *p* | Estimate (*df*) | *p* | Estimate (*df*) | *p* |
| **Movement direction** | $\chi^2(2) = 161$ | < 0.001 | $\chi^2(2) = 62$ | < 0.001 | $\chi^2(2) = 1.47$ | 0.48 |
| **Visualization intensity** | $\chi^2(3) = 11.6$ | 0.009 | $\chi^2(3) = 3.4$ | 0.34 | $\chi^2(3) = 5.7$ | 0.13 |
| **Task phase** | $\chi^2(1) = 110$ | < 0.001 | $\chi^2(1) = 18.4$ | < 0.001 | $\chi^2(1) = 619$ | < 0.001 |
| **Stimulus modality** | $\chi^2(1) = 1530$ | < 0.001 | $\chi^2(1) = 641$ | < 0.001 | $\chi^2(1) = 116$ | < 0.001 |
| **Trial dur. (z-scored)** | M = 0.13 | < 0.001 | M = 0.10 | < 0.001 | M = 0.01 | 0.06 |

## Appendix B3: Task phase comparison for Experiment 1, Hypothesis 2

Model syntax: *Saccade rate ~ movement direction + visualization intensity + trial duration (z-scored) + trial number + modality + (1|subject)*

| **Exp. 1 – Hypothesis 2: Saccade rate [encoding phase only]** | | | |
|---|---|---|---|
| | Estimate | CI | p |
| **(Intercept)** | -0.05 | -0.16 – 0.06 | 0.34 |
| **Movement direction [vertical]** | 0.03 | -0.00 – 0.07 | 0.09 |
| **Movement direction [horizontal]** | 0.03 | -0.01 – 0.06 | 0.14 |
| **Visualization intensity [5]** | -0.22 | -0.39 – -0.05 | 0.01 |
| **Visualization intensity [3]** | -0.19 | -0.39 – 0.02 | 0.07 |
| **Visualization intensity [2]** | 0.11 | -0.14 – 0.37 | 0.38 |
| **Trial duration (z-scored)** | 0.02 | 0.01 – 0.04 | 0.002 |
| **Trial number** | 0.00 | 0.00 – 0.00 | <0.001 |
| **Random effects and model fit** | | | |
| **Resid. variance (SD2)** | 0.19 | | |
| **Subject-level intercept** | 0.05 | | |
| **ICC** | 0.22 | | |
| **Observations** | 3412 | | |
| **R2 (marg./cond.)** | 0.07 / 0.27 | | |

| **Wald test results: Saccade rate in encoding and recall phases** | | | | |
|---|---|---|---|---|
| | **Encoding** | | **Recall** | |
| | Estimate | p | Estimate | p |
| **Movement direction** | $\chi^2(2) = 3.5$ | 0.18 | $\chi^2(2) = 1.2$ | 0.56 |
| **Visualization intensity** | $\chi^2(3) = 10$ | 0.002 | $\chi^2(3) = 2.8$ | 0.42 |
| **Stimulus modality** | $\chi^2(1) = 41.1$ | < 0.001 | $\chi^2(1) = 89$ | < 0.001 |
| **Trial dur. (z-scored)** | M = 0.02 | 0.002 | M = -0.009 | 0.357 |
| **Trial number** | M = 0.001 | < 0.001 | M = 0.002 | < 0.001 |

## Appendix B4: Post-hoc model for Experiment 1, Hypothesis 3

Model syntax: *Travel distance(x) ~ Task phase x Saccade rate + Visualization intensity + (1|Subject)*

| **Exp. 1 – Hypothesis 3: Task phase x Saccade rate** | | | |
|---|---|---|---|
| | Estimate | CI | p |
| **(Intercept)** | 0.16 | -0.10 – 0.43 | 0.229 |
| **Task phase [recall]** | 0.25 | 0.16 – 0.34 | <0.001 |
| **Saccade rate** | 1.55 | 1.50 – 1.60 | <0.001 |
| **Visualization intensity [5]** | -0.25 | -0.69 – 0.19 | 0.261 |
| **Visualization intensity [3]** | 0.36 | -0.15 – 0.87 | 0.169 |
| **Visualization intensity [2]** | 0.63 | -0.01 – 1.27 | 0.053 |
| **Task phase [recall] x Saccade rate** | -0.35 | -0.41 – -0.28 | <0.001 |
| **Random effects and model fit** | | | |
| **Resid. variance ($SD^2$)** | 0.64 | | |
| **Subject-level intercept** | 0.35 | | |
| **ICC** | 0.35 | | |
| **Observations** | 6063 | | |
| **R2 (marg./cond.)** | 0.49 / 0.67 | | |

## Appendix B5: Full model summary (Experiment 2, Hypothesis 4)

| | Travel distance(x) | | | Travel distance(y) | | | Saccade rate | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimate | CI | p | Estimate | CI | p | Estimate | CI | p |
| **(Intercept)** | 1.82 | 1.47 – 2.16 | <0.001 | 2.48 | 2.18 – 2.77 | <0.001 | -0.17 | -0.30 – -0.04 | 0.01 |
| **Motion interpretation [vertical]** | 0.04 | -0.01 – 0.09 | 0.12 | 0.02 | -0.04 – 0.08 | 0.45 | 0.02 | -0.01 – 0.05 | 0.18 |
| **Motion interpretation [horizontal]** | 0.02 | -0.03 – 0.06 | 0.53 | -0.01 | -0.07 – 0.04 | 0.68 | 0.03 | 0.00 – 0.06 | 0.02 |
| **Visualization intensity [4]** | 0.12 | -0.30 – 0.53 | 0.59 | 0.03 | -0.31 – 0.37 | 0.87 | -0.02 | -0.20 – 0.17 | 0.86 |
| **Visualization intensity [3]** | -0.37 | -0.84 – 0.11 | 0.13 | -0.12 | -0.50 – 0.27 | 0.56 | 0.02 | -0.19 – 0.23 | 0.85 |
| **Visualization intensity [2]** | 0.18 | -0.54 – 0.89 | 0.62 | 0.33 | -0.25 – 0.91 | 0.27 | -0.02 | -0.34 – 0.30 | 0.90 |
| **Epoch duration (z-scored)** | 1.91 | 1.81 – 2.01 | <0.001 | 1.80 | 1.68 – 1.92 | <0.001 | -0.44 | -0.49 – -0.39 | <0.001 |
| **Epoch [pre-button]** | 0.09 | 0.01 – 0.17 | 0.02 | 0.21 | 0.12 – 0.30 | <0.001 | 0.18 | 0.13 – 0.22 | <0.001 |
| **Epoch [pre-voice]** | 0.17 | 0.10 – 0.25 | <0.001 | 0.47 | 0.38 – 0.56 | <0.001 | 0.15 | 0.11 – 0.19 | <0.001 |
| **Epoch [articulation]** | -1.92 | -2.10 – -1.75 | <0.001 | -1.63 | -1.84 – -1.42 | <0.001 | 0.72 | 0.62 – 0.82 | <0.001 |
| **Random effects and model fit** | | | | | | | | | |
| **Resid. variance ($\sigma^2$)** | 0.64 | | | 0.89 | | | 0.20 | | |
| **Subject-level intercept** | 2.34 | | | 1.03 | | | 0.06 | | |
| **Subject-level slope for rate** | 0.80 | | | 0.75 | | | – | | |
| **Correlation subject-rate** | -0.93 | | | -0.89 | | | – | | |
| **ICC** | 0.79 | | | 0.54 | | | 0.24 | | |
| **$R^2$ (marg./cond.)** | 0.26 / 0.84 | | | 0.32 / 0.69 | | | 0.14 / 0.35 | | |

## Appendix B5: Full model summary (Experiment 2, Hypothesis 4)

### Wald-test results

| | Travel distance(x) | | Travel distance(y) | | Saccade rate | |
|---|---|---|---|---|---|---|
| | Estimate (*df*) | *p* | Estimate (*df*) | *p* | Estimate (*df*) | *p* |
| **Motion interpretation** | $\chi^2(2) = 2.5$ | 0.28 | $\chi^2(2) = 1.2$ | 0.55 | $\chi^2(2) = 5.3$ | 0.07 |
| **Visualization intensity** | $\chi^2(3) = 4.3$ | 0.23 | $\chi^2(3) = 2$ | 0.57 | $\chi^2(3) = 0.1$ | 0.99 |
| **Epoch** | $\chi^2(3) = 509$ | < 0.001 | $\chi^2(3) = 245$ | < 0.001 | $\chi^2(3) = 572$ | < 0.001 |
| **Epoch dur. (z-scored)** | M = 1.9 | < 0.001 | M = 1.8 | < 0.001 | M = -0.4 | < 0.001 |

## Appendix B6: Post-hoc epoch comparisons (Experiment 2, Hypothesis 4)

Model syntax: *Travel distance(x) ~ Motion interpretation + Visualization intensity + Epoch duration (z-scored) + Epoch + (1|Subject), data filtered for epochs = "audio_replay" & "articulation"*

| **Exp. 2, Hypothesis 4: Travel distance (x) in audio replay / articulation** | | | |
|---|---|---|---|
| | *Estimate* | *CI* | *p* |
| **(Intercept)** | 1.40 | 1.13 – 1.68 | <0.001 |
| **Movement interpretation [vertical]** | 0.08 | 0.03 – 0.13 | 0.001 |
| **Movement interpretation [horizontal]** | 0.04 | -0.01 – 0.08 | 0.13 |
| **Visualization intensity [2]** | 0.12 | -0.27 – 0.51 | 0.54 |
| **Visualization intensity [3]** | -0.42 | -0.87 – 0.03 | 0.07 |
| **Visualization intensity [4]** | 0.31 | -0.36 – 0.99 | 0.36 |
| **Epoch duration (z-scored)** | 1.50 | 1.40 – 1.61 | <0.001 |
| **Epoch [articulation]** | -1.17 | -1.36 – -0.98 | <0.001 |
| **Random effects and model fit** | | | |
| **Resid. variance ($SD^2$)** | 0.39 | | |
| **Subject-level intercept** | 2.09 | | |
| **Subject-level slope for rate** | 2.91 | | |
| **Correlation subject-rate** | -0.93 | | |
| **ICC** | 0.84 | | |
| **Observations** | 4354 | | |
| **R2 (marg./cond.)** | 0.21 / 0.88 | | |

| | **Wald-test results: Travel distance (x) in epoch comparison** | | | |
|---|---|---|---|---|
| | **audio replay & articulation** | | **pre-button & pre-voice** | |
| | Estimate (*df*) | *p* | Estimate (*df*) | *p* |
| **Motion interpretation** | $\chi^2(2) = 10.5$ | 0.005 | $\chi^2(2) = 2.1$ | 0.36 |
| **Visualization intensity** | $\chi^2(3) = 6.9$ | 0.08 | $\chi^2(3) = 3.6$ | 0.31 |
| **Epoch** | $\chi^2(3) = 776$ | < 0.001 | $\chi^2(3) = 591$ | < 0.001 |
| **Epoch dur. (z-scored)** | M = 1.5 | < 0.001 | M = 2.0 | < 0.001 |

## Appendix B7: Full model summary (Experiment 2, Hypothesis 5)

| | Travel distance(x) | | | Travel distance(y) | | | Saccade rate | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimate | CI | p | Estimate | CI | p | Estimate | CI | p |
| (Intercept) | 1.44 | 1.25 – 1.63 | <0.001 | 0.10 | -0.06 – 0.26 | 0.22 | 0.06 | -0.01 – 0.13 | 0.10 |
| Epoch [pre-button] | -0.51 | -0.58 – -0.44 | <0.001 | -0.37 | -0.46 – -0.29 | <0.001 | 0.33 | 0.30 – 0.37 | <0.001 |
| Epoch [pre-voice] | -0.72 | -0.78 – -0.66 | <0.001 | -0.38 | -0.45 – -0.31 | <0.001 | 0.38 | 0.34 – 0.41 | <0.001 |
| Epoch duration (z-scored) | 0.49 | 0.46 – 0.51 | <0.001 | 0.45 | 0.42 – 0.49 | <0.001 | -0.13 | -0.14 – -0.11 | <0.001 |
| Motion interpretation [vertical] | -0.01 | -0.08 – 0.05 | 0.71 | -0.01 | -0.08 – 0.07 | 0.9 | 0.01 | -0.03 – 0.05 | 0.56 |
| Motion interpretation [horizontal] | -0.04 | -0.11 – 0.02 | 0.18 | -0.03 | -0.10 – 0.05 | 0.48 | 0.01 | -0.03 – 0.05 | 0.57 |
| **Random effects and model fit** | | | | | | | | | |
| Resid. variance ($\sigma^2$) | 0.67 | | | 0.97 | | | 0.22 | | |
| Subject-level intercept | 1.6 | | | 1.00 | | | 0.05 | | |
| Subject-level slope for rate | 0.49 | | | 0.50 | | | – | | |
| Correlation subject-rate | -0.88 | | | -0.87 | | | – | | |
| ICC | 0.71 | | | 0.51 | | | 0.18 | | |
| $R^2$ (marg./cond.) | 0.13 / 0.74 | | | 0.11 / 0.56 | | | 0.15 / 0.31 | | |

## Appendix B7: Full model summary (Experiment 2, Hypothesis 5)

### Wald-test results

| | **Travel distance(x)** | | **Travel distance(y)** | | **Saccade rate** | |
|---|---|---|---|---|---|---|
| | Estimate (*df*) | *p* | Estimate (*df*) | *p* | Estimate (*df*) | *p* |
| **Epoch** | $\chi^2(2) = 560$ | < 0.001 | $\chi^2(2) = 128$ | < 0.001 | $\chi^2(2) = 586$ | < 0.001 |
| **Motion interpretation** | $\chi^2(2) = 1.88$ | 0.39 | $\chi^2(2) = 0.53$ | 0.77 | $\chi^2(2) = 0.46$ | 0.79 |

## Appendix B8: Post-hoc model for Experiment 2, Hypothesis 5

Model syntax: *Travel distance(x) ~ Epoch x Saccade rate + Epoch duration (z-scored) + motion interpretation + (1|Subject)*

| **Travel distance (x) by Epoch x Saccade Rate** | | | |
|---|---|---|---|
| | *Estimate* | *CI* | *p* |
| **(Intercept)** | -0.32 | -0.51 – -0.13 | 0.001 |
| **Epoch [pre-button]** | 0.99 | 0.87 – 1.10 | <0.001 |
| **Epoch [pre-voice]** | 0.23 | 0.10 – 0.36 | <0.001 |
| **Saccade rate** | 1.25 | 1.18 – 1.31 | <0.001 |
| **Epoch duration (z-scored)** | 0.44 | 0.42 – 0.47 | <0.001 |
| **Motion interpretation [vertical]** | 0.02 | -0.04 – 0.08 | 0.58 |
| **Motion interpretation [horizontal]** | 0.00 | -0.06 – 0.06 | 0.92 |
| **Epoch [pre-button] x Saccade rate** | -1.06 | -1.13 – -0.99 | <0.001 |
| **Epoch [pre-voice] x Saccade rate** | -0.75 | -0.83 – -0.67 | <0.001 |
| **Random effects and model fit** | | | |
| **Resid. variance ($SD^2$)** | 0.62 | | |
| **Subject-level intercept** | 0.32 | | |
| **ICC** | 0.34 | | |
| **Observations** | 4093 | | |
| **$R^2$ (marg./cond.)** | 0.34 / 0.56 | | |

## Appendix B9: Full model summary (Experiment 2, Hypothesis 6)

| | Travel distance(x) | | | Travel distance(y) | | | Saccade rate | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimate | CI | p | Estimate | CI | p | Estimate | CI | p |
| (Intercept) | 1.38 | 1.08 – 1.67 | <0.001 | -0.05 | -0.32 – 0.22 | 0.72 | 0.01 | -0.13 – 0.16 | 0.88 |
| Explicit constituent [ground] | -0.26 | -0.92 – 0.39 | 0.43 | 0.26 | -0.53 – 1.05 | 0.52 | -0.14 | -0.53 – 0.25 | 0.48 |
| Explicit constituent [direction] | -0.01 | -0.10 – 0.09 | 0.90 | -0.00 | -0.12 – 0.11 | 0.94 | 0.00 | -0.06 – 0.06 | 0.96 |
| Explicit constituent [ground & direction] | 0.07 | -0.03 – 0.17 | 0.15 | -0.07 | -0.18 – 0.05 | 0.25 | 0.03 | -0.03 – 0.09 | 0.31 |
| Motion interpretation [vertical] | 0.02 | -0.05 – 0.09 | 0.62 | 0.02 | -0.06 – 0.11 | 0.56 | -0.00 | -0.05 – 0.04 | 0.86 |
| Visualization intensity [2] | 0.17 | -0.25 – 0.59 | 0.43 | 0.04 | -0.35 – 0.42 | 0.86 | 0.01 | -0.18 – 0.20 | 0.90 |
| Visualization intensity [3] | -0.34 | -0.82 – 0.14 | 0.17 | -0.06 | -0.50 – 0.38 | 0.78 | 0.08 | -0.13 – 0.29 | 0.47 |
| Visualization intensity [4] | 0.18 | -0.54 – 0.90 | 0.63 | 0.41 | -0.25 – 1.07 | 0.22 | 0.07 | -0.23 – 0.38 | 0.65 |
| Epoch [pre-button] | -0.49 | -0.58 – -0.40 | <0.001 | -0.34 | -0.45 – -0.23 | <0.001 | 0.35 | 0.29 – 0.40 | <0.001 |
| Epoch [pre-voice] | -0.70 | -0.78 – -0.62 | <0.001 | -0.33 | -0.43 – -0.23 | <0.001 | 0.35 | 0.30 – 0.40 | <0.001 |
| Epoch duration (z-scored) | 0.47 | 0.43 – 0.51 | <0.001 | 0.43 | 0.39 – 0.48 | <0.001 | -0.11 | -0.14 – -0.09 | <0.001 |
| Random effects and model fit | | | | | | | | | |
| Resid. variance ($\sigma^2$) | 0.64 | | | 0.93 | | | 0.22 | | |
| Subject-level intercept | 1.84 | | | 1.11 | | | 0.05 | | |
| Subject-level slope for rate | 0.65 | | | 0.68 | | | – | | |
| Correlation subject-rate | -0.91 | | | -0.87 | | | – | | |
| ICC | 0.74 | | | 0.55 | | | 0.19 | | |
| $R^2$(marg./cond.) | 0.14 / 0.78 | | | 0.11 / 0.60 | | | 0.16 / 0.32 | | |

## Appendix B9: Full model summary (Experiment 2, Hypothesis 6)

### Wald-test results

| | Travel distance(x) | | Travel distance(y) | | Saccade rate | |
|---|---|---|---|---|---|---|
| | Estimate (*df*) | *p* | Estimate (*df*) | *p* | Estimate (*df*) | *p* |
| Explicit constituent | $\chi^2(3) = 2.9$ | 0.41 | $\chi^2(3) = 1.8$ | 0.61 | $\chi^2(3) = 1.6$ | 0.65 |
| Motion interpretation | $\chi^2(1) = 0.24$ | 0.62 | $\chi^2(1) = 0.34$ | 0.56 | $\chi^2(1) = 0.03$ | 0.86 |
| Visualization intensity | $\chi^2(3) = 4.5$ | 0.21 | $\chi^2(3) = 1.82$ | 0.61 | $\chi^2(3) = 0.71$ | 0.87 |
| Epoch | $\chi^2(2) = 307$ | < 0.001 | $\chi^2(2) = 58$ | < 0.001 | $\chi^2(2) = 247$ | < 0.001 |

## Appendix B10: Removing *epoch duration* as a control variable (Exp. 2, Hyp. 6)

| Travel distance(x) | | | |
|---|---|---|---|
| | *Estimate* | *CI* | *p* |
| **(Intercept)** | 1.32 | 1.03 – 1.62 | <0.001 |
| **Explicit constituent [ground]** | -0.06 | -0.80 – 0.68 | 0.87 |
| **Explicit constituent [direction]** | 0.04 | -0.07 – 0.15 | 0.47 |
| **Explicit constituent [ground & direction]** | 0.13 | 0.02 – 0.24 | 0.02 |
| **Motion interpretation [vertical]** | -0.08 | -0.16 – -0.01 | 0.03 |
| **Visualization intensity [2]** | 0.27 | -0.16 – 0.69 | 0.22 |
| **Visualization intensity [3]** | -0.28 | -0.77 – 0.21 | 0.26 |
| **Visualization intensity [4]** | 0.23 | -0.51 – 0.96 | 0.55 |
| **Epoch [pre-button]** | -0.65 | -0.75 – -0.55 | <0.001 |
| **Epoch [pre-voice]** | -0.78 | -0.87 – -0.69 | <0.001 |
| **Random effects and model fit** | | | |
| **Resid. variance ($SD^2$)** | 0.81 | | |
| **Subject-level intercept** | 1.04 | | |
| **Subject-level slope for rate** | 0.44 | | |
| **Correlation subject-rate** | -0.82 | | |
| **ICC** | 0.56 | | |
| **$R^2$ (marg./cond.)** | 0.09 / 0.60 | | |