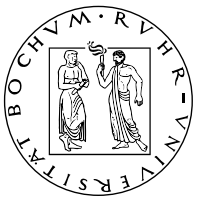


Ruhr Universität Bochum
Institut für Neuroinformatik

Masterarbeit zur Erlangung des akademischen Grades
eines Master of Cognitive Science

A Neural Dynamic Model for the Perceptual Grounding of Combinatorial Concepts

Master Thesis of:
Daniel Sabinasz

First supervisor:
Prof. Gregor Schöner

Second supervisor:
Dr. Mathis Richter

July 2019

Abstract

The ability to combine concepts to model the world in intricate ways underlies most of the higher cognitive competences like thought and language. It has motivated the stance that cognition is purely syntax-driven computation on systems of amodal symbols - a view widely referred to as the *Classical Computational Theory of Mind* (CCTM).

Grounded Cognition (GC) rejects this view and emphasizes the importance of being able to *ground* concepts in perception, i.e., to establish a connection between concepts and objects in the perceptual input.

This master thesis is part of a research program to provide a neural processing account for GC based on neural principles formalized in *Dynamic Field Theory* (DFT). It introduces a neural architecture for the grounding of combinatorial concepts, i.e., concepts that are built by combining other concepts. The architecture receives an arbitrary input image or video and an arbitrary combinatorial concept, which describes an object in terms of its attributes and relationships to other objects – e.g., "a red triangle which is to the right of a red circle that is below a green diagonal rectangle and above a blue object". Its task is to ground the concept in the perceptual input, i.e., to bring the described object into the attentional foreground.

The components of a combinatorial concept are grounded in a sequence of grounding steps, while the output of each grounding step is passed on to the next grounding step through self-sustained neural fields. This way, semantic compositionality is an emergent property of the neural dynamics and does not require any form of amodal symbolic computation.

The capabilities of the architecture are demonstrated in a set of 6 qualitatively different simulations that vary with the complexity of the combinatorial concept and the perceptual input. The architecture is able to successfully ground the given combinatorial concept in all test cases.

Another contribution of this thesis is a clear interface between the grounding system and language, and an embedding in the literature of psychological theories of concepts.

The discussion focuses on motivating aspects of the architecture by theoretical arguments and empirical evidence, contrasting our approach with other accounts for the representation, processing and perceptual grounding of combinatorial concepts, and addressing how our architecture can account for the productivity, compositionality and systematicity of thought and language, which so far have been taken to be the hallmark of CCTM approaches.

Contents

List of Figures

List of Tables

List of Acronyms

ASS Amodal Symbol System 14

AVS Attention Vector Sum Model 52

CCTM Classical Computational Theory of Mind i

CTOC Classical Theory of Concepts 21

DFT Dynamic Field Theory i

DH Dynamical Hypothesis 18

DNF Dynamic Neural Field 32

DNN Dynamic Neural Node 35

DPA Distribution of Population Activation 32

EB Elementary Behavior 42

EC Embodied Cognition 17

GC Grounded Cognition i

GSEnc Grounding Strategy Encoder 6

GSEx Grounding Strategy Executor 6

CoS Condition of Satisfaction 42

LOTH Language of Thought Hypothesis 7

NEF Neural Engineering Framework 116

PA Parallel Architecture 6

PoC Principle of Compositionality 12

PSS Perceptual Symbol System 19

Colophon

This document was typeset using the XeTeX typesetting system created by the Non-Roman Script Initiative and the memoir class[1] created by Peter Wilson. It is based on the PhD thesis template[2] by Frederico Maggi. It is further inspired by the PhD thesis of Mathis Richter (2018). The body text is set 12pt with Adobe Caslon Pro.

[1] http://www.ctan.org/pkg/memoir

[2] https://www.latextemplates.com/template/maggi-memoir-thesis

Introduction 1

The abilities to represent, learn, apply and reason with concepts are prolific features of human cognition and underly most of the higher cognitive functions like thought, belief and language. It is in terms of concepts that we represent the world, reason about the world, and communicate about the world through language. Concepts make up the building blocks of thoughts and beliefs, and most contemporary theories of semantics take them to be the meanings of words or phrases. Understanding the neural basis of concepts is thus essential to understanding the higher cognitive feats and consequently essential for creating general-purpose artificial intelligence.

A noticeable feature of human conceptual systems is the ability to combine concepts, which allows for the construction of a virtually indefinite range of new concepts out of other concepts to model the world in intricate ways. For instance, we can combine the concepts RED, DIAGONAL, and RECTANGLE into the combinatorial concept of a RED DIAGONAL RECTANGLE. Similarly, we can combine the concepts RED, OBJECT, ABOVE, and GREEN into the combinatorial concept of a RED OBJECT ABOVE A GREEN OBJECT. It is this cognitive feat that underlies the production and understanding of nested linguistic expressions, thoughts, and beliefs.

The ability to productively combine concepts has motivated the stance that higher cognitive functions like thought, language understanding and language production are best understood as computation carried out on abstract systems

of symbols – a view that is now widely referred to as the *Classical Computational Theory of Mind* (CCTM) (Rescorla, 2017). In that view, conceptual knowledge resides in symbolic memory systems that are separate from the brain's modal systems for perception, action and introspection. Representations in these modal systems are transduced into symbolic representations with combinatorial syntactic and semantic structure, and higher cognition is the result of purely syntactical processes operating on these representations. Importantly, these symbolic representations are believed to be *amodal*, i.e.,

> "(1) they are arbitrarily related to their corresponding categories in the world and experience; and (2) they can stand alone without grounding to perform the basic computations underlying conceptual processing" (Barsalou, 2016, p. 1125)

Rescorla, M. (2017). The computational theory of mind. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2017). Metaphysics Research Lab, Stanford University

Barsalou, L. W. (2016). On staying grounded and avoiding quixotic dead ends. *Psychonomic Bulletin & Review*, *23*(4), 1122–1142

For example, the proposition "all humans are mortal" could be encoded as the symbol string

$$\forall x(\textsc{human}(x) \rightarrow \textsc{mortal}(x)). \tag{1.1}$$

Similarly, the proposition "Socrates is a human" could be encoded as the symbol string

$$\textsc{human}(\textsc{socrates}). \tag{1.2}$$

The idea behind the CCTM is that the conclusion that "Socrates is mortal", which human beings draw with ease, is the result of a process that makes reference to the syntactic properties of the symbol strings alone. For instance, such a process could contain a rule that whenever a statement with the syntactic form

$$\forall x(A(x) \rightarrow B(x)) \tag{1.3}$$

for arbitrary predicate symbols A and B is encountered in memory, and another statement with the syntactic form

$$A(t) \tag{1.4}$$

for an arbitrary term symbol t is encountered in addition, then the system should produce the statement

$$B(t). \tag{1.5}$$

This rule would lead the system to the conclusion

$$\text{MORTAL}(\text{SOCRATES}) \tag{1.6}$$

from the premises given by Equation 1.1 and Equation 1.2, which states that "Socrates is mortal".

Defenders of this view contend that only the CCTM can account for the ability to productively combine concepts, the fact that language and concepts feature a compositional semantics, and the fact that the ability to understand some concepts is systematically related to the ability to understand certain other concepts (Fodor & Pylyshyn, 1988). Consequently, many computational models of language processing and conceptual reasoning are given in terms of algorithmic symbol processing.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71

In the last few decades, the CCTM has been challenged by various findings and arguments, e.g., the lack of empirical evidence for amodal symbols in the brain, overlap in the brain regions responsible for perception and conceptual reasoning, the symbol grounding problem, and inconsistencies with neural principles of computation (see Section 2.4).

In response, views of *Grounded Cognition* (GC) reject that conceptual knowledge is stored and processed in symbolic memory stores. Instead, they propose that higher cognitive processes rely on neural representations stored and processed in the same neural systems that are also responsible for perception and whose representational format mirrors the perceptual states that produced them. According to this view, conceptual systems pervasively rely on multimodal simulations and situated action, and adequate models of conceptual systems cannot abstract away from these aspects, as the CCTM attempts (see Section 2.5).

Essential to GC is the ability to establish a connection between concepts and the perceptual array or the world – a process referred to as *grounding* (Gorniak & Roy, 2004). As Harnad (1990) has argued, the mere processing of abstract symbols devoid of meaning makes it impossible to actually attain an understanding of what the symbols are about. When we draw the conclusion that "Socrates is mortal" from the premises "all humans are mortal" and "Socrates is a human", then we have a sense that we actually understand this reasoning. The conclusion does not just pop out of nowhere, as would be the case if it were produced by an amodal symbol processing algorithm. Proponents of GC

Gorniak, P. & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*, 335–346

propose that this sense of understanding is the result of being able to ground these statements and their constituents in perception. When hearing statements like "all humans are mortal", we can ground this state of affairs, e.g., by imagining it visually. Thus, GC emphasizes that at any point during conceptual processing, the cognitive system has an ability to attain an understanding of what this processing is about by grounding the statements, concepts, and constituents of the concepts in perception.

GC is a verbal theory rather than a concrete model. This master thesis is part of a research program to provide a neural processing account for GC. The account is based on *Dynamic Field Theory* (DFT), a mathematical framework for the modeling of cognitive processes that is rooted in our understanding of neural principles. Previous work by the DFT research community has, amongst other things, provided neural architectures for perception, detection, selection, attention, scene representation, sequence generation, process organization, and motor control (Schöner & Spencer, 2015). Moreover, neural architectures have been devised for the grounding of spatial or movement relations between two objects (Lipinski et al., 2012; Richter et al., 2014; Richter et al., 2017; Richter, 2018). They are able to ground denotational phrases like "a red object below a blue object" or "a red object that is moving towards a blue object" – or, equivalently, the combinatorial concepts that are the meanings of these linguistic expressions, which are provided by the user in the form of an activation pattern over feature concept nodes and relation concept nodes. These architectures demonstrate how the discrete symbolic representations that underlie language can be linked to continuous perceptual representations close to the sensorimotor layer. Moreover, they demonstrate how properties and systematic relations among multiple objects can be recognized and expressed, which is a fundamental aspect of human intelligence in general and language in particular. However, these architectures are limited to a single relation between two objects specified by a single attribute value.

The primary goal of this master thesis is to build upon this work and devise a neural architecture that allows for the grounding of arbitrarily nested combinatorial concepts, e.g., "a red triangle which is to the right of a red circle that is below a green diagonal rectangle and above a blue object".

Schöner, G. & Spencer, J. (2015). *Dynamic thinking: A primer on dynamic field theory.* New York, NY: Oxford University Press

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neuro-behavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *38*(6), 1490–1511

Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., & Schöner, G. (2014). Autonomous neural dynamics to test hypotheses in a model of spatial language. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 2847–2852). Austin, TX: Cognitive Science Society

Richter, M., Lins, J., & Schöner, G. (2017). A neural dynamic model generates descriptions of object-oriented actions. *Topics in Cognitive Science*, *9*, 35–47

Richter, M. (2018). *A neural dynamic model for the perceptual grounding of spatial and movement relations.* (Doctoral dissertation, Bochum, Ruhr-Universität Bochum)

The architecture receives sensory input from a camera or an image file, which is used to feed a continuous model of the environment. Furthermore, it receives a grounding strategy in the form of a neurally encoded sequence of parameterized instructions that have to be performed in order to ground a given combinatorial concept, e.g., "find all green objects, eliminate all non-diagonal objects, eliminate all non-rectangles, make a selection decision, find a blue object, ...". The model then performs this sequence of instructions in order to ground the concept, i.e., to find an object in the scene that matches this concept. For the purpose of illustration, the model is restricted to low-level perceptual attributes like color, orientation, shape, and to spatial relations. However, the architecture is modular, so that it can be extended to arbitrary other perceptual or conceptual spaces. Thus, the architecture serves as a blueprint for a generic architecture that is able to ground arbitrary combinatorial concepts in perception.

In building this grounding system, particular emphasis is put on the following aspects:

neural principles of computation — All building blocks of the architecture cohere to established neural principles of computation. This involves not only sticking to computations that can be performed by networks of neurons instead of introducing algorithmic building blocks, but also to make sure that the way the neurons interact is biologically plausible. This distinguishes the architecture from many other computational models, which often employ algorithmic techniques that are impossible to be performed by the neural hardware of the brain.

autonomy — The architecture unfolds solely based on the initial sensory input, grounding strategy input, and its internal dynamics. In particular, it does not require any external control input by a human user.

emergent discrete events — Discrete events like processing steps, detection or selection decisions emerge from the internal dynamics through dynamic instabilities.

stability — The components of the architecture form stable representations that are robust to changing sensory

input and noise.

coordination of shared resources — Different processes are coordinated and do not interfere with each other. At any time, the state of the architecture coheres with the demands of the currently active processes.

A second goal of this master thesis is to embed the neural theory of atomic and combinatorial concepts underlying our grounding system in the literature of theories of concepts. It will be argued that the DFT-based accounts of atomic concepts, which have already been prevalent in past architectures, are prototype theories of concepts in the form of a probability distribution in conceptual or perceptual spaces. Moreover, it will be argued that the class of combinatorial concepts that our new architecture is able to ground are all those concepts that can be described as a graph consisting of *frames* and relations between frames. A frame consists of a set of attribute-value assignments. Frame graphs are a common concept representation format that is widely believed to be able to capture a wide range of human concepts, and that is adopted with slight variations in both amodal and perceptual theories of knowledge, as well as natural language semantics (e.g., Minsky, 1977; Sowa et al., 2000; Barsalou, 1999; Jackendoff, 2002).

A third goal is to clearly interface the grounding system with language. In doing so, I draw upon the *Parallel Architecture* (PA) developed by Jackendoff (2002), according to which language has multiple parallel sources of combinatoriality. Auditory input is analyzed for phonological structure, which is analyzed for syntactical structure, which is analyzed for conceptual structure. According to our idea, conceptual structure is then transformed into a sequential grounding strategy by a *Grounding Strategy Encoder* (GSEnc). This grounding strategy is in turn executed by the *Grounding Strategy Executor* (GSEx), effectively grounding the original linguistic phrase in perception and thereby facilitating actual understanding of the linguistic phrase.

A fourth goal is to demonstrate the capabilities of the grounding system in a set of simulations. Test cases differ in the perceptual inputs and to-be-grounded combinatorial concepts, which have been chosen to demonstrate qualitatively different scenarios of varying complexity. In all

Minsky, M. (1977). A framework for representing knowledge. In P. Winston (Ed.), *The psychology of computer vision*. New York, NY: McGraw-Hill

Sowa, J. F. et al. (2000). *Knowledge representation: Logical, philosophical, and computational foundations*. Pacific Grove, CA: Brooks/Cole

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press

simulations, the architecture is able to successfully ground the given combinatorial concepts, and the components of the architecture form stable representations that cohere to the demands of active processes.

A fifth goal is to motivate some of the design features of the architecture by theoretical arguments and empirical evidence.

A sixth goal is to contrast our approach with other neural or CCTM models of the representation, processing or grounding of combinatorial concepts. It will be argued that our architecture exhibits productivity, compositionality, and systematicity, which so far have been the hallmark of CCTM approaches. It will be demonstrated that they emerge from the perceptual grounding process due to the order in which a sequence of grounding steps is carried out, and due to the way the activation state representing the perceptual grounding of one object is carried over to the grounding of a subsequent object. This deprives proponents of the CCTM of one of their main arguments and is a step towards demonstrating that neural processing accounts for GC can serve as fully functional conceptual systems, which are able to account for the same range of cognitive capacities that have previously been argued to lend indirect support to the CCTM.

All in all, this master thesis brings DFT closer toward a neural theory of conceptual processing. By extension, it brings DFT closer toward a neural theory of language understanding, language production, and ultimately cognition as a whole.

The remainder of this thesis is structured as follows. Chapter 2 provides background knowledge that is required to understand the architecture and to embed it into a broader research context. It summarizes the CCTM, the *Language of Thought Hypothesis* (LOTH), and the main arguments given in favor of these views, followed by a review of the challenges to these theories. It goes on to describe the GC research program, theories of atomic and combinatorial concepts, and the *Parallel Architecture* (PA) of language processing. It then reviews the mathematical and conceptual foundations of DFT. Chapter 3 introduces the Grounding Strategy Encoder, a brain system that takes a representation of a combinatorial concept as input and transforms it into a sequence of parameterized instructions that have to be performed in order to ground

that concept. Chapter 4 gives a description and mathematical formalization of the Grounding Strategy Executor, which takes the grounding strategy as input and performs it. Chapter 5 gives results and explanations of a range of experiments conducted with the architecture. These experiments demonstrate the various concept-grounding capabilities of the architecture. Chapter 6 features a discussion of the architecture, contrasts it with other accounts for the representation or perceptual grounding of combinatorial concepts, and considers the implications of this work for the debate between the CCTM and GC. Chapter 7 concludes the thesis.

Background 2

This chapter provides the background knowledge required to understand the grounding system for combinatorial concepts, to embed it into its research context, and to pave the way for the discussion that contrasts our approach with other approaches to the representation and apprehension of combinatorial concepts.

Section 2.1 summarizes the *Classical Computational Theory of Mind* (CCTM), the view that a brain is nature's way of implementing a computer. Section 2.2 describes the *Language of Thought Hypothesis* (LOTH), a species of the CCTM emphasizing the representational, combinatorial and compositional character of symbols, which has strongly established itself in computational modeling of conceptual processing and in linguistics. Section 2.3 reconstructs a prominent set of arguments given in favor of the CCTM and LOTH by Fodor and Pylyshyn (1988). Section 2.4 reviews a range of empirical findings and arguments that challenge the CCTM and LOTH. Section 2.5 introduces the *Grounded Cognition* (GC) view, which has been put forward in response to these challenges and highlights the perceptual nature of the higher cognitive competences. Section 2.6 summarizes influential psychological theories of concepts. Section 2.7 describes the *Parallel Architecture* (PA), an influential theory of the relationship between combinatorial concepts and language. Section 2.8 describes the grounding process in some detail. Finally, Section 2.9 reviews *Dynamic Field Theory* (DFT), paving the way for the model of the grounding of combinatorial concepts.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71

2.1 The Classical Computational Theory of Mind

As already alluded to in the introduction, a historical trend in the cognitive sciences has been to understand the brain as nature's way of implementing a computer, a view often termed the *Classical Computational Theory of Mind* (CCTM). On that view, sensory input gets transformed into a symbolic representational format, and higher cognition is the result of purely syntactical operations carried out on these symbolic representations, i.e., mechanistic operations that only make reference to the physical shape of the symbols, not to their meanings (Rescorla, 2017).

Rescorla, M. (2017). The computational theory of mind. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2017). Metaphysics Research Lab, Stanford University

Early roots of this idea can be found in the development of formal logics as a means for modeling laws of reason. A formal logic is a formal language consisting of a set of symbols, rules for combining these symbols into complex symbol structures, and purely syntactical rules of inference, i.e., rules that only make reference to the form of the symbols rather than their meaning. Through the application of these syntactical rules, true conclusions can be derived from true premises. This way, the syntax "tracks" the semantics: While the rules of transformation are purely syntactical, they are defined in such a way that truth is preserved across syntactical transformations – i.e., that syntactical operations satisfy semantical coherence.

For example, a syntactical rule could be specified that allows to derive the string "The sky is blue" from the string "The sky is blue and the sun is shining" due to the fact that the former is a constituent of the latter. This tracks the semantic fact that the proposition denoted by "The sky is blue" follows from the proposition denoted by "The sky is blue and the sun is shining". This inference is valid, and it can be made without reference to the meaning of either "The sky is blue" or "The sun is shining". As such, formal logics allow to account for reasoning processes without intrinsic reference to meaning.

Formal logics introduce a liberalism regarding the actual physical shapes of the symbols, as long as the shapes are used consistently, and the rules of transformation are set up in accordance with these shapes. Thus, instead of using the symbol "and" for conjunction, we could use the symbol "#", so long as we formulate the rules of inference in accordance,

e.g., that the strings "A" and "B" can be derived from the string "A # B".

Another milestone in the development of the CCTM was the formalization of the notion of computation by Turing (1936) in the form of the *Turing machine*, a hypothetical device that is able to execute any algorithm and solve any decidable problem. Importantly, its operations consist of the transformation of symbols, and these operations are sensitive to the syntactical structure (and only the syntactical structure) of those symbols. In combination with the work on formal logics, this showed that it is possible to construct an autonomous, syntax-driven machine whose state transitions satisfy semantical coherence – i.e., a reason-respecting machine.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*(2), 230–265

McCulloch and Pitts (1943) were among the first to suggest that the human mind is nature's way of implementing something that is similar in important respects to a Turing machine. The primary motivation for this view is its ability to account for the mental in naturalistic terms, and for how reason-respecting behavior can emerge from the interaction of physical matter. In the course of the 1960s, this stance was at the heart of the emerging field of cognitive science. It was widely believed that many of the higher cognitive functions like reasoning, decision making and problem solving are computations carried out in a fashion similar to a Turing machine. Research in mathematical modeling of cognition was thus closely intertwined with the emerging fields of computer science and artificial intelligence, and models of cognitive processes were often given in algorithmic terms.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133

2.2 The Language of Thought Hypothesis

While the core thesis of the CCTM is that the brain implements a Turing-style computational mechanism, many species of the CCTM include additional tenets. A common additional tenet is that the symbols manipulated in the computations are mental representations (Pitt, 2018). A mental representation is a structure with semantic properties – it represents something in the environment. Depending on one's theory of mental representations, they may represent objects, categories of objects, properties, relations, states

Pitt, D. (2018). Mental representation. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2018). Metaphysics Research Lab, Stanford University

of affairs, or any combination of them. This tenet is often referred to as the *Representational Theory of Mind* (RTM). The fact that symbols have meanings is usually formalized by a *meaning function* μ that maps symbols to meanings (Werning, 2005). For example, the symbol SWAN may represent the set of all swans in the world, i.e.,

Werning, M. (2005). Right and wrong reasons for compositionality. *The Compositionality of Meaning and Content: Foundational Issues*, *1*, 285–309

$$\mu(\text{SWAN}) = \{x | x \text{ is a swan}\}. \tag{2.1}$$

Another common additional tenet is that the mental representations manipulated in mental computations have a part/whole constituency structure, i.e., that there are syntactical operations that allow to combine symbols into complex symbols, which can be combined into yet more complex symbols, etc. This is often referred to by saying that the symbols themselves are "combinatorial" (Fodor & Pylyshyn, 1988).

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71

The syntactical operations that allow to combine symbols into complex symbol structures are usually formalized as functions $\sigma : S^n \to S$ from a sequence of n symbols into the set of symbols. For instance, there might be a syntactical rule that allows to combine adjective and noun symbols into a noun phrase symbol, which can be formalized as a function

$$\begin{aligned} \sigma_{\text{AdjectiveNounCombination}} &: \text{A} \times \text{N} \to \text{NP}, \\ \sigma_{\text{AdjectiveNounCombination}}(a, n) &= a\ n \end{aligned} \tag{2.2}$$

from the Cartesian product of the set of all adjectives and the set of all nouns to the set of all noun phrases. This syntactical rule allows, e.g., to combine the symbols BLACK and SWAN into the combinatorial symbol BLACK SWAN:

$$\sigma_{\text{AdjectiveNounCombination}}(\text{BLACK}, \text{SWAN}) = \text{BLACK SWAN} \tag{2.3}$$

A third additional tenet, which goes hand in hand with the ability to combine symbols into combinatorial symbols, is the *Principle of Compositionality* (PoC), according to which the meaning of a combinatorial symbol structure is determined by the meanings of its parts and the way the parts are put together syntactically (Janssen et al., 2012):

Janssen, T. M. et al. (2012). Compositionality: Its historic context. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The oxford handbook of compositionality* (pp. 19–46)

Definition 2.2.1 (Compositional meaning function) *A meaning function μ is called* compositional *if for every syntactic operation σ, there is a function μ_σ such that*

$$\mu(\sigma(s_1, \ldots, s_n)) = \mu_\sigma(\mu(s_1), \ldots, \mu(s_n)). \tag{2.4}$$

For example, if the meanings of adjectives and nouns are sets of objects in the world to which the adjective or noun applies, then the meaning of a noun phrase may be given by the intersection of the set to which the adjective applies with the set to which the noun applies:

$$\mu_{\sigma_{\text{AdjectiveNounCombination}}}(\mu(a), \mu(n)) = \mu(a) \cap \mu(n) \tag{2.5}$$

Thus, if the meaning of the adjective BLACK is the set of all black things, i.e.,

$$\mu(\text{BLACK}) = \{x | x \text{ is black}\}, \tag{2.6}$$

and the meaning of the noun SWAN is the set of all swans, i.e.,

$$\mu(\text{SWAN}) = \{x | x \text{ is a swan}\}, \tag{2.7}$$

then the meaning of the combinatorial symbol BLACK SWAN is the set of all black swans, i.e.,

$$\begin{aligned}\mu(\text{BLACK SWAN}) =& \mu_{\sigma_{\text{AdjectiveNounCombination}}}(\mu(\text{BLACK}), \mu(\text{SWAN})) \\ =& \mu(\text{BLACK}) \cap \mu(\text{SWAN}) \\ =& \{x | x \text{ is black}\} \cap \{x | x \text{ is a swan}\} \\ =& \{x | x \text{ is black and } x \text{ is a swan}\}.\end{aligned} \tag{2.8}$$

Some authors speak of "compositional symbol structures" or "compositional concepts" to refer to what is meant here by a "combinatorial symbol structure" or a "combinatorial concept". This can lead to confusion with the PoC, which makes a claim about the semantics of the symbol structures, not merely about the fact that they can be combined.

The three additional tenets that (1) symbols are mental representations with (2) a part/whole constituency structure that (3) fulfill the PoC are at the heart of the *Language of Thought Hypothesis* (LOTH), a species of the CCTM put forward by Fodor (1975). An important feature of the LOTH is that the systems responsible for the higher cognitive functions are encapsulated modules that are independent from and cannot penetrate the sensory-motor systems (Fodor, 1983). Rather, perceptual representations are transformed into a completely new representational format, which is inherently symbolic and amodal. These amodal representations are fed into the systems responsible for higher cognition, and their output can be used to program motor commands.

Fodor, J. A. (1975). *The language of thought.* New York, NY: Crowell

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press

One of the major attractions of the LOTH and derived views has been their ability to serve as fully functional conceptual systems that are able to account for many of the higher cognitive feats like memory, knowledge, reasoning, language, and thought.

The systems of symbols on which the brain is believed to operate under the CCTM and LOTH are often referred to as *Amodal Symbol Systems* (ASSs) to highlight two aspects of this view: (1) The relation between the symbols and their meaning is arbitrary, i.e., the symbols attain their meaning by convention, or by their role in the overall symbol system, but do not themselves resemble the objects in the world or the perceptual states that produced them. (2) The symbols enter into conceptual processing without grounding them in perception (Barsalou, 1999).

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609

There is a slight confusion in the literature regarding the use of the term "amodal". On some occasions, it is taken to refer to symbols that fulfill both (1) and (2). On other occasions, it is taken to refer to symbols that only fulfill (1), i.e., symbols that are arbitrarily related to what they denote but may or may not require grounding to enter into conceptual processing. To avoid confusion, I shall only use the term to refer to structures that fulfill both (1) and (2). For structures that are only arbitrarily related to what they denote but may or may not require grounding, I shall use the term "non-perceptual symbol". Thus, non-perceptual symbols may be grounded in perception, whereas amodal symbols cannot be grounded by definition. Examples for non-perceptual symbols include words or abstract concepts. While these symbols might not be represented in the brain's modal systems, they are not necessarily amodal, provided that they can be grounded, i.e., that a connection can be established between these symbols and modal representations.

2.3 The productivity, compositionality, and systematicity challenge

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71

In an influential rebuttal of connectionism, Fodor and Pylyshyn (1988) argue that the productivity, compositionality and systematicity of language and thought can be

accounted for by the CCTM and, as they conjecture, only the CCTM. Since the model presented in this thesis rejects the CCTM, it is of value to look at these arguments in some detail. Later, we will see how these arguments can be circumvented.

productivity — The mind is *productive* in the sense that it can produce and understand an indefinite range of thoughts, concepts and linguistic expressions by finite means. This, Fodor and Pylyshyn argue, requires that these representations belong to a set that is generated recursively out of parts – something that a Turing machine is able to do.

systematicity — The *systematicity* of thought and language refers to the fact that the capacity to produce and understand some thoughts or linguistic expressions is systematically related to the ability to produce and understand certain others. For example, the ability to understand "Alice sees Bob" is systematically related to the ability to understand "Bob sees Alice". In particular, there can be no cognitive system which understands the one but not the other. Thus, the argument goes, there must be structural relations between the thoughts/linguistic expressions. Combinatorial structure of thought and language, as proposed by the LOTH, predicts systematicity.

compositionality — Fodor and Pylyshyn claim that natural language meets the PoC. Since language is used to express thoughts, they infer that thoughts must also meet the PoC. This requires that thoughts have internal structure. Since thoughts do have internal structure in the LOTH, it can account for compositionality.

systematicity of inference — Lastly, Fodor and Pylyshyn argue that inferences of a similar logical type should involve similar cognitive capacities. For instance, an inference from $P\&Q\&R$ to P should be executed by the same computational process as an inference from $P\&Q$ to P. This is a strong argument against early connectionist models, which modeled the processing of combinatorially structured representations as the

spreading of activation from nodes representing combinatorial structures to nodes representing their constituents. For instance, an inference from $P\&Q\&R$ to P was modeled by spreading activation from a neuron representing $P\&Q\&R$ to a neuron representing P, while an inference from $P\&Q$ to P was modeled in a separate process by the spreading of activation from a neuron representing $P\&Q$ to a neuron representing P. Fodor and Pylyshyn claim that this is implausible, since it allows for gaps in one's reasoning capacities, e.g., being able to infer that "Alice arrived" from the proposition "Alice and Bob and Charlie arrived", but not being able to infer it from the proposition "Alice and Bob arrived".

The commonality in all these arguments is that they show that the mind must somehow operate on structured representations, that its operations must be sensitive to the combinatorial structure of those representations, and that there must be a homogeneous mechanism that processes combinatorial structures, regardless of what the parts of those structures are and how they are combined.

An important aspect of these properties is that while the CCTM can account for them and classical connectionist architectures cannot[3], they do not show that *only* the CCTM can account for them. As will be discussed in Section 6.3, our architecture can address these challenges while being significantly less powerful than a Turing machine, thus being more neurally plausible.

[3] When introducing recurrency into connectionist architectures, they begin to be able to account for these properties. Fodor and Pylyshyn might respond that in this case, the connectionist network only serves as an implementation of a CCTM architecture and thus, at a cognitive level, is more adequately modeled as a CCTM architecture. However, recurrent connectionist architectures can be built that are significantly less powerful than a Turing machine and still meet these challenges. This has to date not been addressed by proponents of the CCTM.

2.4 Challenges to the Classical Computational Theory of Mind

In the last few decades, the CCTM been challenged by various empirical findings and arguments, many of which are reviewed by Barsalou (1999, 2008). For instance, there is no direct empirical evidence for amodal symbols in the brain. To the contrary, picture naming studies suggest that conceptual symbols have a perceptual nature (Glaser, 1992). In line with this, language processing studies suggest that sentence meaning composition depends on deriving

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645

Glaser, W. R. (1992). Picture naming. *Cognition*, *42*(1-3), 61–105

affordances from sensory-motor simulations (Glenberg & Robertson, 2000).

Neuroscientific evidence suggests that brain regions responsible for perception also become active during conceptual reasoning (Pulvermüller, 2005) and that damage to a perceptual region impairs conceptual reasoning about object classes whose perceptual processing uses that region (Pulvermüller, 1999). This speaks against a divide between a perceptual domain on the one hand and a syntactical computational domain on the other hand.

Furthermore, the CCTM is faced with the symbol grounding problem (Harnad, 1990), i.e., the lack of a satisfactory account for how abstract symbols can get mapped back to perceptual representations and objects in the world, and for how a sense of understanding of one's reasoning can come about in the absence of referents for the symbols. Harnad demonstrates this by a thought experiment, according to which a person has to learn Chinese using only a Chinese-Chinese dictionary. Since this dictionary only explains some meaningless symbols in terms of other meaningless symbols, the argument goes, the person cannot possibly gain an understanding of Chinese. Harnad likens this situation to an amodal symbol processing system: Since the system only blindly manipulates amodal symbols, it cannot possibly gain a sense of understanding what these symbols mean.

Related to the symbol grounding problem is the converse problem that there is no satisfactory account for how a perceptual state can get mapped to an abstract symbol in the first place, and no neuroscientific evidence that such a process exists.

Another line of criticism comes from the *Embodied Cognition* (EC) research program. This program is quite heterogeneous and there is no generally agreed-upon definition for it. Most proponents emphasize how the body, the environment, situated action, and their dynamic coupling shape cognition in complex ways that do not get accounted for in detached symbol manipulation (e.g., Shapiro, 2010). Additionally, some of these theories emphasize the close coupling of perception and action during goal achievement, which speaks against a divide between perception, discrete symbol processing and action (e.g., Clark, 1997). As a result, they emphasize that realistic models of cognitive processes should be situated in a real environment.

In line with this view, many authors (e.g., Van Gelder,

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

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(July), 576–582

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, *22*(2), 253–336

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*, 335–346

Shapiro, L. (2010). *Embodied cognition.* Routledge

Clark, A. (1997). *Being there: Putting brain, body, and world together again.* Cambridge, MA: MIT Press

1998) have suggested what is now widely referred to as the *Dynamical Hypothesis* (DH), according to which cognitive systems are dynamical systems in closed loop with their environment and are best modeled as such: sensory input affects the internal dynamics of a cognitive system, which shapes the motor output, which in turn may change the sensory input. Usually, dynamical systems models reject the idea of fixed representations and a central processor. Instead, cognition is taken to emerge from coupled specialized systems, each of which can reside in one of infinitely many continuous states. Additional support for the dynamical hypothesis comes from the fact that cognition unfolds in continuous time, whereas Turing-style computation unfolds in discrete processing stages (Port & Van Gelder, 1995).

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and brain sciences*, *21*(5), 615–628

Port, R. F. & Van Gelder, T. (1995). *Mind as motion: Explorations in the dynamics of cognition.* Cambridge, MA: MIT press

Computational models of cognitive functions put forward by proponents of the CCTM are often inconsistent with neural principles of computation (e.g., Richter et al., 2017). The kinds of computations that are known to be supported by the brain are significantly more restricted than the computations supported by a Turing machine.

Richter, M., Lins, J., & Schöner, G. (2017). A neural dynamic model generates descriptions of object-oriented actions. *Topics in Cognitive Science*, *9*, 35–47

On a related note, neural models attempting to account for combinatorial structure in representations usually suggest that neural activation patterns can encode high-dimensional representations through mostly orthogonal activity patterns in which each activity value stands for another dimension, suggested, e.g., by Smolensky (1990). In addition to the lack of stability that such architectures are faced with, they would require an encoding from sensory information to high-dimensional activity vectors and a subsequent decoding from these high-dimensional vectors to motor commands. This, as argued by Schöner and Spencer (2015, p. 86), makes it difficult to maintain a smooth coupling between sensory input and motor action.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1-2), 159–217

2.5 Grounded Cognition

While the EC research program emphasizes the role of the body and situated action for cognition, the GC research program additionally emphasizes that cognition is inherently perceptual, i.e., that there is no qualitative division between cognitive processes at the sensory-motor level and higher cognitive processes (Barsalou, 2008). Instead, these processes only differ in their distance to the sensory-motor

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645

surfaces, but follow the same neural principles.

An aspect of this view is that most higher cognitive processes rely on multimodal simulations: the brain captures multimodal perceptual, motor, and introspective states during experience with the world and integrates them into multimodal representations in long-term memory. When an object of a certain category is encountered, or knowledge about a category is needed, multimodal representations of instances of that category are partially reactivated to simulate how the brain represented those instances while perceiving them or interacting with them.

Most proponents of GC reject that amodal symbols exist in the brain. Those who leave it open whether they exist postulate that they work alongside modal representations.

Barsalou (1999) proposes that higher cognitive processes rely on *Perceptual Symbol Systems* (PSSs). According to this view, subsets of the perceptual states that arise in sensory-motor systems are extracted through selective attention and stored in long-term memory in the form of *perceptual symbols*. Importantly, perceptual symbols can later be retrieved and function symbolically, i.e., stand for objects or states of affairs in the world and enter into symbol manipulation processes. Thereby, they give rise to the higher cognitive competences.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609

Perceptual symbols are modal in the sense that they are stored and processed in the same perceptual brain systems as the states that gave rise to them. Moreover, their internal structures resemble the perceptual states that produced them to a certain degree. This is in contrast to amodal symbols, which are believed to be stored and processed in separate brain systems, and which bear an arbitrary relation to what they denote.

The perceptual symbols extracted from perception and interaction with an object of a given category can be integrated into a *simulator* for that category. This simulator is composed of perceptual symbols that have been extracted from previous category members. Upon each new encounter with a category member, the simulator becomes extended by more multimodal information of what it is like to perceive, think about, and interact with members of that category. Over time, the simulator thus accumulates a large amount of multimodal information allowing to simulate instances of that category. Simulators can develop not only for object categories, but also for relations. Moreover, simulators can

be combined into combinatorial simulators. For instance, Barsalou demonstrates how simulators for the categories *balloon*, *jet*, *cloud*, *above*, and *left of* can be combined into a simulator for the category *a ballon that is above a jet to the left of a cloud*.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609

Barsalou (1999) states that the CCTM and LOTH are justified in emphasizing the significance of symbolic operations for the emergence of higher cognitive functions. In particular, he grants that one of the merits of ASSs is their ability to function as a fully functional conceptual system, which

> "[...] represents both types and tokens, it produces categorical inferences, it combines symbols productively to produce limitless conceptual structures, it produces propositions by binding types to tokens, and it represents abstract concepts." (Barsalou, 1999, p. 581)

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71

Thus, his view is not directly opposed to the arguments presented by Fodor and Pylyshyn (1988). However, Barsalou demonstrates that they can be met by PSSs as opposed to ASSs. In the course of his article, he develops a theory that hints at how PSSs may serve as fully functional conceptual systems.

Barsalou's theory is a verbal theory rather than a concrete model. The work on the neural basis of GC by the DFT research community, including the architecture proposed in this master thesis, aspire to provide neural process models for Barsalou's theory.

2.6 Concepts

In psychology, the term *concept* is commonly defined as follows:

> "A concept of x is a body of knowledge about x that is stored in long-term memory and that is used by default in the processes underlying most, if not all, higher cognitive competences when these processes result in judgments about x." (Machery, 2009, p. 11)

Machery, E. (2009). *Doing without concepts*. Oxford University Press

Thus, concepts are mental representations used by the higher cognitive competences. Opinions differ as to what

these higher cognitive competences are, but most commonly they are taken to include categorization, induction, reasoning, analogy-making, conceptual combination, language production and language understanding. In cognitive linguistics, concepts are usually taken to be the meanings/referents of words. Many philosophers of mind also take concepts to be the building blocks of thoughts (Margolis & Laurence, 2019).

Margolis, E. & Laurence, S. (2019). Concepts. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2019). Metaphysics Research Lab, Stanford University

Before the 1960s, the prevalent view in the philosophy of mind and in psychology was that a concept of a category of objects is a set of separately necessary and jointly sufficient conditions for category membership (Smith & Medin, 1981). This view, now often called the *Classical Theory of Concepts* (CTOC), interfaced neatly with the CCTM and LOTH, and many computational models of concept acquisition and logical reasoning embraced it.

Smith, E. E. & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press

There is now considerable empirical evidence that the CTOC is inadequate. First, most human concepts do not have sharp boundaries, but are vague (e.g., Keefe & Smith, 1996). Rather than judging whether or not a given concept is applicable, humans instead seem to determine a degree of membership. For instance, robins are commonly judged to be more typical instances of birds than penguins (Hampton, 2007), which is taken to show that humans assign robins a higher degree of membership in the category of birds than penguins.

Keefe, R. & Smith, P. (1996). *Vagueness: A reader*. Cambridge, MA: MIT Press

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science*, *31*(3), 355–384

Moreover, Machery (2009) reviews evidence which suggests that the class of concepts divides into three heterogeneous kinds, namely, *prototypes*, *exemplars* and *theories*.

Machery, E. (2009). *Doing without concepts*. Oxford University Press

Prototypes are mental representations of average or prototypical instances of a category. For example, a prototype of my category of dogs may be a prototypical or average dog representation. Prototype theories differ in how these prototypes are represented, but common formats are attribute lists (Rosch & Mervis, 1975), regions or probability distributions in perceptual attribute spaces or conceptual spaces (Gärdenfors, 2000), and frames (see next section; Smith, 1988). As an example, consider the concept RED. Gärdenfors (2014) proposes that this concept is represented in the brain by a prototypical instance of red in HSV color space, and that the degree of membership of a color

Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605

Smith, E. E. (1988). Concepts and thought. *The Psychology of Human Thought*, *147*

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge, MA: MIT Press

in the category red is determined by the similarity between the prototype and the color instance.

Exemplars are mental representations of concrete instances of a category. For instance, a memory of my dog Teddy is an exemplar for my concept of dog. Different theories propose different representational formats for exemplars, but commonly use points in perceptual attribute spaces, points in conceptual spaces (Gärdenfors, 2000), or frames (Nosofsky, 1986). Exemplar theories of concepts propose that one's concept of a category is the sum total of one's exemplars of that category.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39

Theories are bodies of propositional knowledge about a category, and are usually introduced by analogy with scientific theories. The idea is that humans maintain a body of propositional knowledge about a category as part of their concept of that category. This factual knowledge about a category may be embedded within a more encompassing theory about a wider subject matter. For instance, my concept of dog may include the propositional knowledge that dogs coevolved with humans, which is embedded within a wider theory of biological species and their evolution.

While prototype theories, exemplar theories and theory theories have initially been proposed as competing theories of how humans represent concepts, reviews of empirical evidence suggest that humans possess all three types of concepts (Murphy, 2002; Machery, 2009).

Murphy, G. (2002). *The big book of concepts.* Cambridge, MA: MIT Press

Machery, E. (2009). *Doing without concepts.* Oxford University Press

A noticeable feature of human conceptual systems is the ability to combine concepts, which allows for the construction of a virtually indefinite range of new concepts out of other concepts to model the world in intricate ways (e.g., Hampton & Winter, 2017). In the following, I use the term *combinatorial concept* to refer to concepts that are built by combining other concepts, and the term *atomic concept* to refer to concepts that cannot be further subdivided into building blocks.

Hampton, J. A. & Winter, Y. (2017). *Compositionality and concepts in linguistics and psychology.* Springer International Publishing

2.6.1 Atomic concepts

Atomic concepts constitute the basic building blocks of the human conceptual system. They can be classified into

attributes, values, and relations (Barsalou, 1992).

Barsalou, L. W. (1992). Frames, concepts, and conceptual fields. In A. Lehrer & E. F. Kittay (Eds.), *Frames, fields, and contrasts: New essays in lexical and semantic organization.* Lawrence Erlbaum Associates, Inc

Attributes are concepts that describe an aspect of an object. Examples include COLOR, ORIENTATION, SHAPE, SIZE, or LOCATION. Attributes can be categorical (e.g., shape, which may take the categorical values RECTANGLE, CIRCLE, etc.) or quantitative. Quantitative attributes can be discrete (e.g., number of legs) or continuous (e.g., size). Barsalou (1992) notes that concepts are only attributes when they describe an aspect of a larger whole. Thus, for instance, color becomes an attribute when viewed as an aspect of a bird, but is not an attribute when viewed in isolation. Note that the term *feature* is sometimes used instead of *attribute*, especially in DFT models. To remain consistent with naming conventions from both frame theory and DFT, I use both terms interchangeably.

Values are subordinate concepts of attributes. For instance, RED is a value of COLOR, DIAGONAL is a value of ORIENTATION and TRIANGLE is a value of SHAPE. In exemplar models, attributes are usually assigned a single value (e.g., 45° for ORIENTATION). In prototype models, they can be assigned a range of values or a set of values to model the fact that the category encompasses multiple possible values as opposed to a single value. It is also common to assign them a probability distribution over values to model different degrees of membership. For example, the concept DIAGONAL could be modeled as a Gaussian centered on 45°. Orientations close to 45° would then have a high degree of membership, whereas orientations far from 45° would have a low degree of membership.

Relations are also concepts, but they are special in that they do not describe categories or aspects of objects, but rather relationships between objects.

In this master thesis, we restrict ourselves to the attributes COLOR with the possible values RED, GREEN, BLUE, YELLOW, ORIENTATION with the possible values HORIZONTAL, DIAGONAL, VERTICAL and SHAPE with the possible values RECTANGLE, SQUARE, ELLIPSE, CIRCLE, TRIANGLE. Furthermore, we restrict ourselves to the spatial relations LEFT OF, RIGHT OF, ABOVE and BELOW.

2.6.2 Combinatorial concepts

Combinatorial concepts are concepts that are built by combining two or more other concepts, which can be atomic or combinatorial as well (Barsalou, 2017). Whereas atomic concepts are often taken to be the meanings of words, combinatorial concepts are often taken to be the meanings of nested linguistic expressions. The relationship between language and concepts is elaborated in more detail in the next section.

It is widely assumed that combinatorial concepts fulfill the PoC, such that the meaning of a combinatorial concept is determined by the meanings of its parts and the way the parts are put together. In this case, they are often referred to as *compositional concepts*.

Combinatorial concepts can be constructed ad hoc during language processing or goal achievement (Barsalou, 1983) or be stored in long-term memory. For instance, the combinatorial concept of a red object below a green diagonal object is an ad hoc concept that we may construct upon reading and understanding a sentence, albeit the atomic concepts that comprise the building blocks of this concept (RED, OBJECT, BELOW, GREEN, DIAGONAL) are stored in long-term memory. In contrast, the combinatorial concept of an uncle as a brother of a parent is stored in long-term memory.

Frames and *frame graphs* are models of combinatorial concepts (Barsalou, 1992). They are widely believed to be able to model many or all combinatorial concepts present in human thought and are used pervasively in exemplar-based and prototype-based models of concepts (e.g., Nosofsky, 1986; Smith, 1988), in natural language semantics (e.g., Jackendoff, 1987), in cognitive psychology under the name "schemata" (Cohen & Murphy, 1984), in amodal theories of knowledge (Sowa et al., 2000), and in modal theories of knowledge (Barsalou, 1999).

A *frame* is a set of attributes that are shared by category members. For example, a frame for the concept CAR could include the attributes MARQUE, COLOR, ENGINE, MAXIMUM SPEED, etc. (Figure 2.1a). Note that this frame-based model of the concept CAR is highly impoverished and only serves for illustrational purposes. More exhaustive models, still based on frames, represent a car as composed of a hierarchy of parts, along with their meronomic relations,

Barsalou, L. W. (2017). Cognitively plausible theories of concept composition. In J. A. Hampton & Y. Winter (Eds.), *Compositionality and concepts in linguistics and psychology* (pp. 9–30). Springer International Publishing

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211–227

Barsalou, L. W. (1992). Frames, concepts, and conceptual fields. In A. Lehrer & E. F. Kittay (Eds.), *Frames, fields, and contrasts: New essays in lexical and semantic organization*. Lawrence Erlbaum Associates, Inc

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39

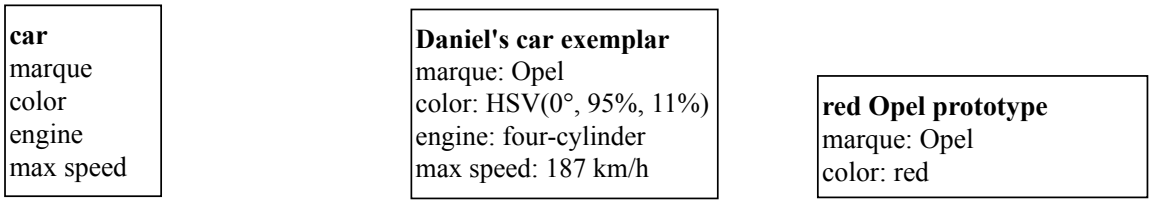
Smith, E. E. (1988). Concepts and thought. *The Psychology of Human Thought*, *147*

Jackendoff, R. (1987). The status of thematic relations in linguistic theory. *Linguistic Inquiry*, *18*(3), 369–411

Cohen, B. & Murphy, G. L. (1984). Models of concepts. *Cognitive Science*, *8*(1), 27–58

Sowa, J. F. et al. (2000). *Knowledge representation: Logical, philosophical, and computational foundations*. Pacific Grove, CA: Brooks/Cole

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609



(a) A car frame. (b) A car exemplar. (c) A red Opel prototype.

FIGURE 2.1: Examples for a car frame. These examples show how the car frame can be applied to model car exemplars and car prototypes.

spatial arrangement and functional relationships (Barsalou, 1999).

Exemplars of a given category assign values to each attribute. For instance, my car could be formalized as the exemplar depicted in Figure 2.1b. Prototypes of a given category may assign values, ranges of values or a probability distribution over values to a subset of the attributes. For instance, a prototype for the concept of a red Opel car could be formalized as the prototype depicted in Figure 2.1c. Note that in contrast to the exemplar, the color attribute is assigned a value of RED instead of a concrete HSV color value. RED may stand for a prototype of the concept of red, e.g., in the form of a Gaussian centered on an average red value.

Introducing relations between frames allows to encode the fact that objects bear certain relations to each other, which allows to express a wider range of propositions and combinatorial concepts. Graphs consisting of frames and relations between frames are commonly referred to as *frame graphs* (Petersen, 2007). Frame graphs are able to represent adjective-noun combinations (e.g., RED APPLE), adjective conjunctions (e.g., RED DIAGONAL OBJECT), noun-noun combinations (e.g., TREE HOUSE), arbitrary denotational phrases, and arbitrary propositions expressible in first-order logic. As such, they promise a unifying role in models of the representational vehicles underlying the higher cognitive functions.

As an example, consider the proposition expressed by the sentence "A red opel crashed into Daniel's car and a blue car." The meaning of that sentence can be regarded as a combinatorial concept, which can be modeled as the frame graph depicted in Figure 2.2. Note that "Daniel's

Petersen, W. (2007). Representation of concepts as frames. *The Baltic International Yearbook of Cognition, Logic and Communication*, *2*, 151–170

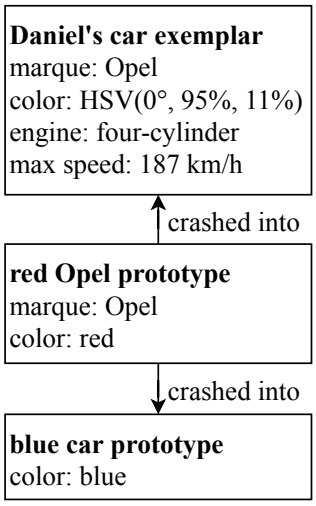


FIGURE 2.2: Frame graph for the proposition "A red opel crashed into Daniel's car and a blue car".

car" is a proper name referring to a concrete car and is therefore modeled as an exemplar, whereas "a red Opel" and "a blue car" are indefinite descriptions (Ludlow, 2018) specified only by their attributes and are therefore modeled as prototypes.

Ludlow, P. (2018). Descriptions. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Fall 2018). Metaphysics Research Lab, Stanford University

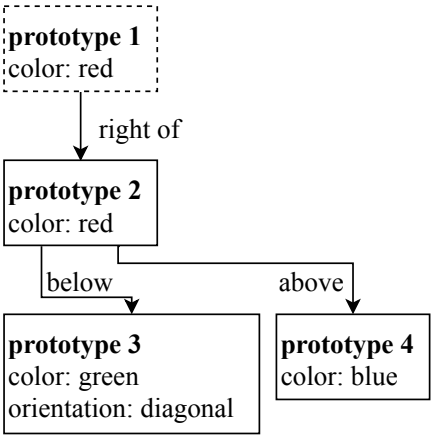


FIGURE 2.3: Combinatorial concept for the noun phrase "a red object right of a red object below a green diagonal object and above a blue object".

In addition to full propositions, frame graphs can also model the meanings of noun phrases. For instance, the noun phrase "a red object right of a red object below a green diagonal object and above a blue object" can be expressed as the frame graph in Figure 2.3. The target of the noun phrase is surrounded by a dashed line.

The notion of a frame graph introduced here is largely equivalent to the notion of a simulator employed by Barsalou. A distinguishing characteristic is that Barsalou's simulators may contain information about the spatial extents and spatial positions of their components, whereas the frame graphs introduced here are reduced to content information. However, it is easy to imagine how to incorporate spatial extents and positions into frame graphs in the form of additional attributes specifying outlines and coordinates.

2.7 The Parallel Architecture

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution.* Oxford University Press

The linguist Jackendoff (2002) develops what he calls the *Parallel Architecture* (PA), a functional description of how language is processed in the human brain. A visualization of this architecture is depicted in Figure 2.4. According to his idea, natural language has multiple parallel sources of combinatoriality, and each of them creates its own characteristic type of structure.

In particular, he proposes that raw speech input is fed into a phonological analysis system, which outputs a phonological structure (e.g., phonetic, syllabic and prosodic structure). This phonological structure is fed into a syntactical analysis system, which generates a syntactical structure (usually, a syntax tree). The syntactical structure is in turn fed into a semantical analysis system that generates a conceptual structure. The conceptual structures generated by his system are equivalent to frame graphs (e.g., Jackendoff, 2002, p. 6).

As an example, consider noun phrases. A noun phrase is a phrase that picks out an object by describing its category, attributes, and relationships with other objects. The

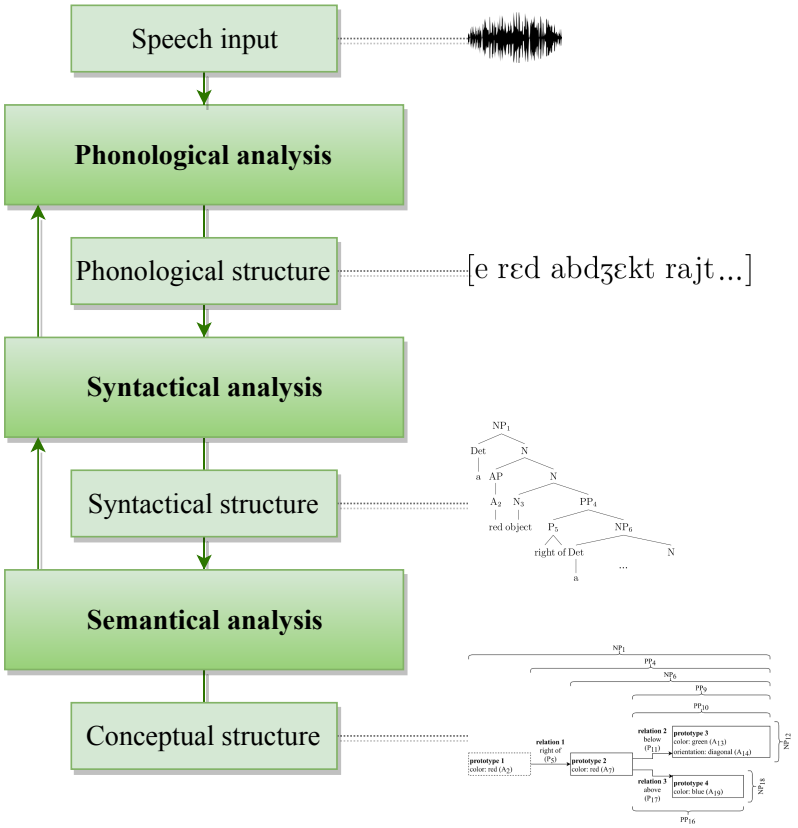


FIGURE 2.4: The Parallel Architecture of language processing developed by Jackendoff (2002). Speech input enters into a phonological analysis system, which outputs a phonological structure (here reduced to a string of phonemes). This phonological structure feeds into a syntactical analysis system, outputting a syntactical structure (here reduced to a syntax tree). The syntactical structure in turn feeds into a semantical analysis system, which constructs a conceptual structure (here a frame graph).

formal grammar depicted in Figure 2.5 allows to generate most of the possible English noun phrases. An exemplary lexicon of nouns, adjectives and prepositions restricted to colors, orientations, shapes, and spatial relations is given in Figure 2.6.

```
1 NP → Det N
2 N → AP N | N PP
3 PP → P NP | PP Conj PP
4 AP → A | AP AP
5 Det → a | the
6 Conj → and | or
```

FIGURE 2.5: Formal grammar for English noun phrases. NP: noun phrase, Det: determiner, N: noun, AP: adjective phrase, PP: prepositional phrase, Conj: conjunction, A: adjective.

```
1 N → object | rectangle | square | ellipse |
    circle | triangle
2 A → red | blue | green | yellow | horizontal |
    diagonal | vertical
3 P → left of | right of | above | below
```

FIGURE 2.6: Formal grammar for an exemplary lexicon of nouns, adjectives and prepositions.

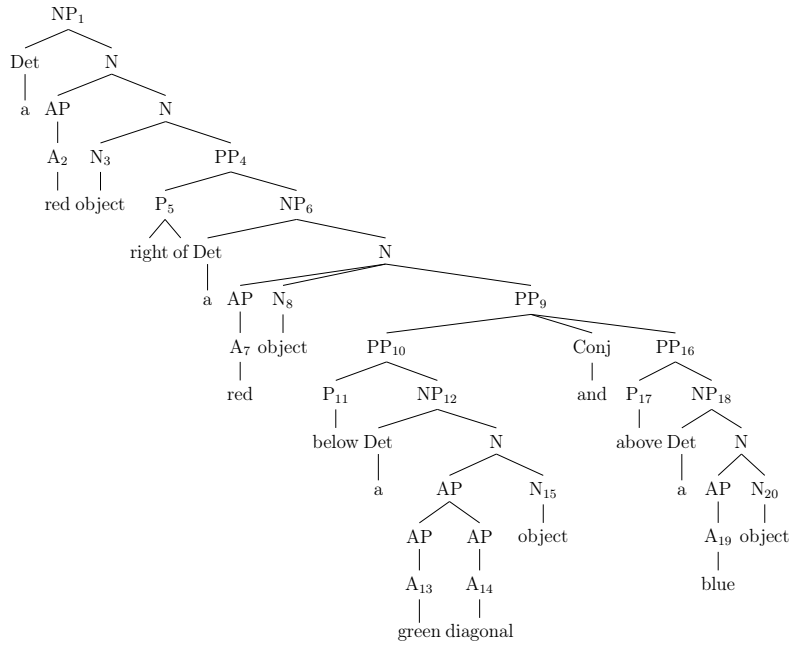


FIGURE 2.7: Syntax tree for the noun phrase "a red object right of a red object below a green diagonal object and above a blue object". Semantically relevant constituents are indexed.

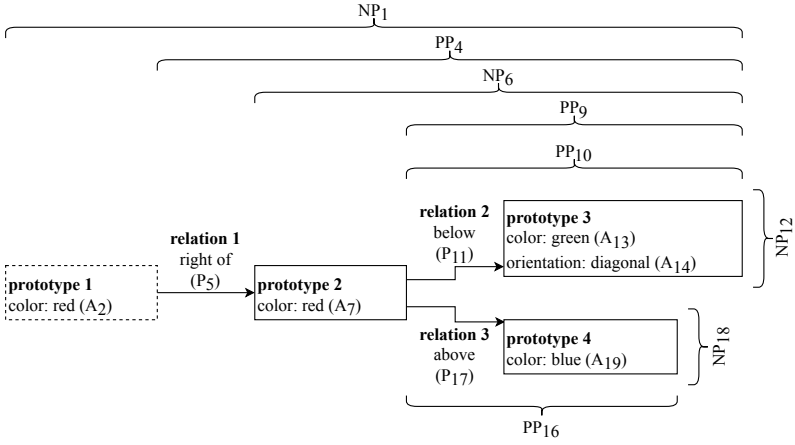


FIGURE 2.8: Frame graph as a conceptual structure for the sentence "a red object right of a red object below a green diagonal object and above a blue object". Constituents are dovetailed by curly brackets and matched to the syntactical constituents from Figure 2.7. Frames and relations are indexed.

An example for a noun phrase that can be generated with that grammar is "a red object right of a red object below a green diagonal object and above a blue object". The PA proposes that speech input is first analyzed for phonological structure, outputting a phonetic representation ("[e rɛd abdʒɛkt rajt əv e rɛd abdʒɛkt bəlo e rin dajæənəl abdʒɛkt ænd əbəv e blu abdʒɛkt]") as well as syllabic, prosodic and morphophonological structure that need not concern us. The phonological structure is analyzed for its syntactic structure in accordance with the formal grammar, outputting the syntax tree depicted in Figure 2.7. This syntax tree is fed into a semantic interpretation system that generates a conceptual structure in the form of the frame graph depicted in Figure 2.8, the conceptual constituents

of which are matched to the syntactical constituents. This conceptual structure can then serve as a basis for other cognitive processes.

The motivation for the PA is primarily linguistic in nature. The fact that phonetic, syntactic and conceptual structure feature the proposed sort of combinatoriality is inferred from observations of natural language behavior; it is implicit in the way we use language. Jackendoff maintains that this combinatoriality is explicitly represented in the brain and not just an emergent property of language, but does not commit to an explicit model for how it is represented. Section 6.2 discusses some neural models for how combinatorial structure may be represented in the brain.

Psycholinguists have long assumed that syntactical and semantic processing proceed in modular systems that are impenetrable by perceptual systems and that perform conceptual processing in the form of amodal symbolic processing. The PA, in particular, is committed to the LOTH. In contrast, GC highlights the importance of grounding language understanding in perception. Thus, we suggest that the PA should be supplemented by a language grounding architecture. This architecture is the focus of this master thesis and will be described in depth in the upcoming chapters.

2.8 The grounding process

One of the claims of grounded cognition is that conceptual processing is inherently perceptual in nature. This requires the ability to establish a connection between concepts and perception, a process we refer to as *grounding*. Two types of grounding processes can be distinguished. First, a concept could be activated by some high-level cognitive process, e.g., by input from a language processing system, and that concept is then grounded by finding and attending to an object in the perceptual array that matches this concept. We refer to that object as the *target* of the grounding process. Second, an already attended object or set of objects could cause a matching concept to be activated or a new combinatorial concept to be built out of other concepts. We refer to the latter process as *describing*. This master thesis focuses on a solution to the first problem.

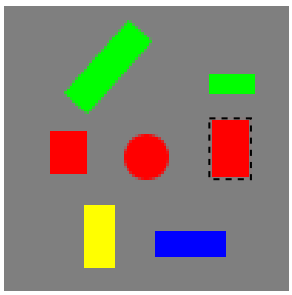


FIGURE 2.9: Example scene. The grounded combinatorial concept is surrounded by dashed lines.

As an example, consider the combinatorial concept describable by the phrase "a red object right of a red object below a green diagonal object and above a blue object" (Figure 2.8). If the perceptual input is given by Figure 2.9, then grounding the concept would surmount to raising attention to the object surrounded by dashed lines.

The nervous system may ground a combinatorial concept via one of three distinct cognitive mechanisms. The first mechanism is a kind of heuristic approach, in which, through repeated acquaintance with objects that fall under a combinatorial concept, characteristic features are learned which allow to identify the instances of that concept without explicitly considering the combinatorial structure of the concept (Keil & Batterman, 1984). For instance, while the concept of a car represents it as composed of a hierarchy of parts, along with their meronomic relations, spatial arrangement and functional relationships, this structure is usually not explicitly considered when identifying cars. Instead, cars are identified by diagnostic features such as their characteristic shape or the sound of a motor.

Keil, F. C. & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 221–236

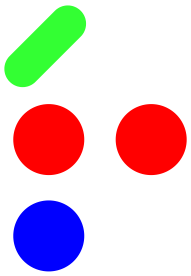


FIGURE 2.10: A mental model for the combinatorial concept "a red object right of a red object below a green diagonal object and above a blue object".

The second mechanism consists in building a mental model of a combinatorial concept, which can be thought of as an image-like representation of a prototypical instance of that concept. For example, upon hearing the phrase "a red object right of a red object below a green diagonal object and above a blue object" (Figure 2.8), a human interpreter is likely to first imagine this arrangement of objects, i.e., to build a mental model (Figure 2.10). This mental model can then guide visual search in the perceptual array, e.g., by trying to match the mental model to a visual arrangement of objects in the perceptual array. A DFT account for building mental models of this kind has been given by Kounatidou et al. (2018). An account for matching mental models to a scene is as of yet lacking.

Kounatidou, P., Richter, M., & Schöner, G. (2018). A neural dynamic architecture that autonomously builds mental models. In C. Kalish, M. A. Rau, X. Zhu, & T. T. Rogers (Eds.), *Proceedings of the 40th annual meeting of the cognitive science society*. Austin, TX: Cognitive Science Society

The third mechanism explicitly considers the structure of the combinatorial concept and matches the parts of the concept (i.e., the frames and relations) to aspects of perception. This strategy is the focus of this master thesis.

2.9 Dynamic Field Theory

Dynamic Field Theory (DFT) (Schöner & Spencer, 2015) is a mathematical and conceptual framework for the mod-

Schöner, G. & Spencer, J. (2015). *Dynamic thinking: A primer on dynamic field theory*. New York, NY: Oxford University Press

eling of cognitive processes. In line with the dynamical hypothesis, DFT models cognitive systems as dynamical systems, i.e., as continuous-time differential equations. In addition, it emphasizes neural principles of computation, embodiment, grounded cognition, and stable states. DFT accounts of many cognitive processes have been given, including perceptual, motor, and grounding processes. These process models are iteratively being joined together into more complex models with the ultimate goal of accounting for cognition as a whole. This section summarizes both the key principles and the mathematical formalization of DFT.

2.9.1 Key principles

Process models — According to the levels of analysis proposed by Marr (1982), DFT models cognitive systems at the *implementational level*, i.e., it models not just what the system does, but also how it is physically realised. In other words, it builds *process models*, which are to be distinguished from algorithmic or statistical models that merely account for what a system does effectively, or how to account for empirical data, but not for how the system does what it does.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: WH Freeman

Neural principles of computation – Most proponents of DFT criticize algorithmic models for being too liberal in the computational operations that they allow for. The computational operations that networks of neurons are able to perform are significantly more restricted than the computational operations that a Turing machine can perform. Thus, a goal of DFT is to provide a coherent framework for building models out of building blocks that are based on established neural principles. These building blocks can be joined together into architectures that comprise process models for cognitive feats.

Population coding — A central tenet of DFT is that behaviorally relevant parameters are coded for by the activity of populations of neurons in circumscribed brain areas, as opposed to the activity of single neurons – an idea referred to as *population coding*. Consequently, all neurons in a given population contribute to behavior, not just, e.g., the most active one. Arguments for this claim include the fact that the activity

of a single neuron is ambiguous with respect to the parameter value that it codes for, and the fact that using just a single neuron to code for a certain parameter value makes behavior highly susceptible to the influence of noise. Moreover, studies have demonstrated population coding for saccade parameters in the superior colliculus (Lee, Rohrer, & Sparks, 1988) and for arm movement parameters in the motor cortex (Georgopoulos, Kettner, & Schwartz, 1988). Consequently, models in DFT do not model individual neurons, as is done, e.g., by connectionists. Instead, the activity of a population of neurons is captured by a *Dynamic Neural Field* (DNF) that is defined over the space of behaviorally relevant parameters to which the population is responsive (e.g., saccade direction and amplitude), assigns a continuous degree of activation to each point in that space, and evolves in continuous time. The activation distribution of a DNF can be regarded as modeling the *Distribution of Population Activation* (DPA) of the neural population that it models. A DPA of a given neural population in a given time interval is obtained by summing up their tuning curves weighted by their average firing rates. The result is then corrected for non-uniform sampling, e.g., by dividing it by the unweighted sum of the tuning curves.

Lee, C., Rohrer, W. H., & Sparks, D. L. (1988). Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature*, *332*(6162), 357

Georgopoulos, A. P., Kettner, R. E., & Schwartz, A. B. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. *Journal of Neuroscience*, *8*(8), 2928–2937

Embodiment — In line with the EC stance, DFT puts emphasis on creating embodied robotic implementations of theoretical models in real environments with real sensory input and a closed sensory-motor loop. First, this can illustrate how the body, the environment, situated action, and their dynamic coupling may shape cognition in intricate ways, which is virtually impossible to model in artificial situations. Second, it can demonstrate that the architecture can function autonomously in the world, as opposed to only in artificial experimental situations. Third, it can demonstrate the robustness of the architecture to real, possibly noisy, sensory inputs. Fourth, it can demonstrate that the dynamical system exhibits stability even if it is in closed loop with the environment, in which case sensory-motor contingencies exist that do not exist in artificial experimental situations and

might lead to unexpected behavior.

Stability — Cognitive systems are subject to many sources of noise. These include the inherent noise in the nervous system, sensor noise and motor noise. Moreover, sensory input is constantly changing. Cognitive systems must resist these sources of noise and maintain invariant representations in the face of them. DFT achieves this through dynamical systems that are in attractor states most of the time, resisting perturbations due to noise and quickly changing environmental circumstances. These attractor states comprise the invariant representations of the system and afford stable behavior.

Instabilities — Dynamic instabilities, which may be induced by significant changes in environmental circumstances or internal representations, can lead to a change in the attractors and, therefore, a change in the system's internal representations. These changes in the attractors occur at discrete moments in time, which can account for how discrete events like detection or selection emerge in continuous time.

2.9.2 Dynamic Neural Fields

In the following, I summarize the mathematical framework underlying DFT. The notational conventions of the equations are mostly adopted from Richter (2018).

Richter, M. (2018). *A neural dynamic model for the perceptual grounding of spatial and movement relations.* (Doctoral dissertation, Bochum, Ruhr-Universität Bochum)

The activity of a population of neurons is captured by a *Dynamic Neural Field* (DNF), which is a function $u(\vec{x}, t)$ defined over the continuous attribute dimensions $\vec{x}$ to which that population is responsive, and over continuous time t. It assigns a continuous degree of activation to each point in that space at each point in time.

The field receives external input from sensors or other fields, formalized as a function $s(x, t)$, and in turn yields an output $g(u(\vec{x}, t))$ that may serve as input to other fields.

$g(u)$ is the sigmoid function (Figure 2.11),

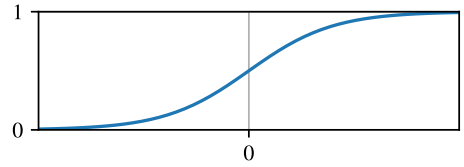


FIGURE 2.11: Sigmoid function $g(u)$.

$$g(u) = \frac{1}{1 + exp(-\beta(u - u_0))}. \tag{2.9}$$

u_0 controls the inflection point of the sigmoid, whereas β controls its steepness. The sigmoid function produces

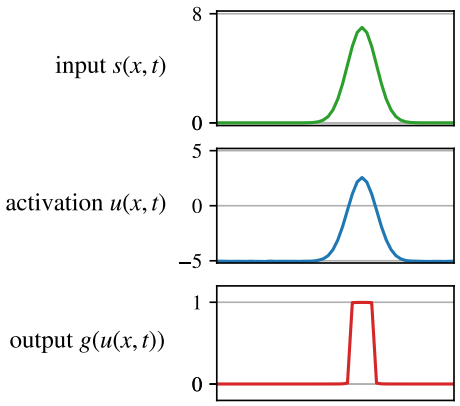


FIGURE 2.12: An exemplary Dynamic Neural Field. From top to bottom: The input $s(x,t)$, the activation $u(x,t)$, and the output $g(u(x,t))$.

Jancke, D., Erlhagen, W., Dinse, H. R., Akhavan, A. C., Giese, M., Steinhage, A., & Schöner, G. (1999). Parametric population representation of retinal location: Neuronal interaction dynamics in cat primary visual cortex. *Journal of Neuroscience*, *19*(20), 9016–9028

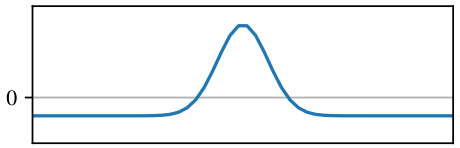


FIGURE 2.13: A lateral interaction kernel with local excitation and global inhibition.

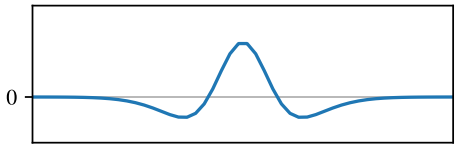


FIGURE 2.14: A lateral interaction kernel with local excitation and mid-range inhibition.

values close to 1 for positive levels of activation, values close to 0 for negative levels of activation, and features a smooth transition in-between. This ensures that only fields that have formed peaks of positive activation have an output that may affect the dynamics of other fields.

Figure 2.12 depicts a snapshot of the input, activation and output of an exemplary 1-dimensional DNF.

The activation $u(\vec{x},t)$ evolves in continuous time t based on the differential equation

$$\begin{aligned}\tau\dot{u}(\vec{x},t) = &-u(\vec{x},t) + h + s(\vec{x},t) + w_\xi \cdot \xi(\vec{x},t) \\ &+ \int g(u(\vec{x}',t))\, k(\vec{x}-\vec{x}')\, d\vec{x}'.\end{aligned} \tag{2.10}$$

The rate of change of the activation depends on a time constant τ, on $u(\vec{x},t)$ itself, on a negative resting level h, on an external input $s(\vec{x},t)$, and on Gaussian white noise $\xi(\vec{x},t)$ with strength w_ξ.

The integral formalizes lateral interactions within the field as a convolution of the output of the field, $g(u(\vec{x}',t))$, with a kernel $k(\Delta\vec{x})$, which will henceforth be abbreviated as $[k * g(u)](\vec{x},t)$. The kernel has a homogeneous structure throughout the field and only depends on the distance $\Delta\vec{x} = \vec{x} - \vec{x}'$ between two points.

In line with findings regarding recurrent interaction patterns in the cortex (Jancke et al., 1999), the kernels chosen by DFT are Mexican-hat functions that feature local excitation and global or mid-range inhibition. Figure 2.13 depicts an exemplary kernel with global inhibition and Figure 2.14 depicts an exemplary kernel with mid-range inhibition. This is formalized by the equation

$$\begin{aligned}k(\Delta\vec{x}) = &\; w_{\text{exc}} \cdot \varphi(\Delta\vec{x}, \vec{\mu}_{\text{exc}}, \vec{\sigma}_{\text{exc}}) \\ &- w_{\text{inh,mid}} \cdot \varphi(\Delta\vec{x}, \vec{\mu}_{\text{inh,mid}}, \vec{\sigma}_{\text{inh,mid}}) \\ &- w_{\text{inh,glob}},\end{aligned} \tag{2.11}$$

where w_{exc}, $\vec{\mu}_{\text{exc}}$ and $\vec{\sigma}_{\text{exc}}$ respectively determine the strength, center and standard deviation of local excitation, $w_{\text{inh,mid}}$, $\vec{\mu}_{\text{inh,mid}}$ and $\vec{\sigma}_{\text{inh,mid}}$ determine the strength, center and standard deviation of mid-range inhibition and $w_{\text{inh,glob}}$ determines the strength of global inhibition. The centers $\vec{\mu}$ are usually set to 0. The strength of local excitation is higher than the strength of mid-range inhibition ($w_{\text{exc}} > w_{\text{inh,mid}}$), whereas the range is lower ($\vec{\sigma}_{\text{exc}} < \vec{\sigma}_{\text{inh,mid}}$). φ is a multi-

variate Gaussian function without covariance defined as

$$\varphi(\Delta\vec{x}, \vec{\mu}, \vec{\sigma}) = a \cdot exp\left(-\sum_{i=1}^{d} \frac{(\Delta\vec{x}_i - \vec{\mu}_i)^2}{2\vec{\sigma_i}^2}\right), \qquad (2.12)$$

where $\vec{\mu}$ is the mean vector, $\vec{\sigma}$ the standard deviation vector and a the amplitude of the Gaussian.

With small or no external input, lateral interactions are not effective, causing the field to remain in a subthreshold attractor at $h + s(\vec{x}, t)$. Due to the term $-u(\vec{x}, t)$ in the field dynamics, the activation is always driven back to this attractor and can thereby resist noise.

When input of sufficient strength is applied, the subthreshold attractor becomes unstable, causing the field to form peaks of positive activation at positions of strong input. The size and number of peaks depend on the parameters of the interaction kernel. With strong global inhibition, a field can form a single peak that inhibits all others. With small or no global inhibition, a field can form multiple peaks, the number of which depends on the range and strength of mid-range inhibition. This is elaborated in more detail in Section 2.9.4.

Peaks are the units of representation in DFT. Through excitatory or inhibitory coupling between fields, the output of one field can serve as input to another field. This way, complex cognitive architectures can be built out of mutually coupled fields.

2.9.3 Dynamic Neural Nodes

Dynamic Neural Nodes (DNNs) follow the same dynamics as fields, but they are 0-dimensional in the sense that they only have a single activation value that evolves in time according to the equation

$$\tau\dot{u}(t) = -u(t) + h + s(t) + w_{\text{se}} \cdot g(u(t)). \qquad (2.13)$$

Instead of a convolution, they have a single weighted self-excitation term with strength w_{se}. These nodes can be coupled to fields and thereby activate continuous representations, as described in Section 2.9.5.

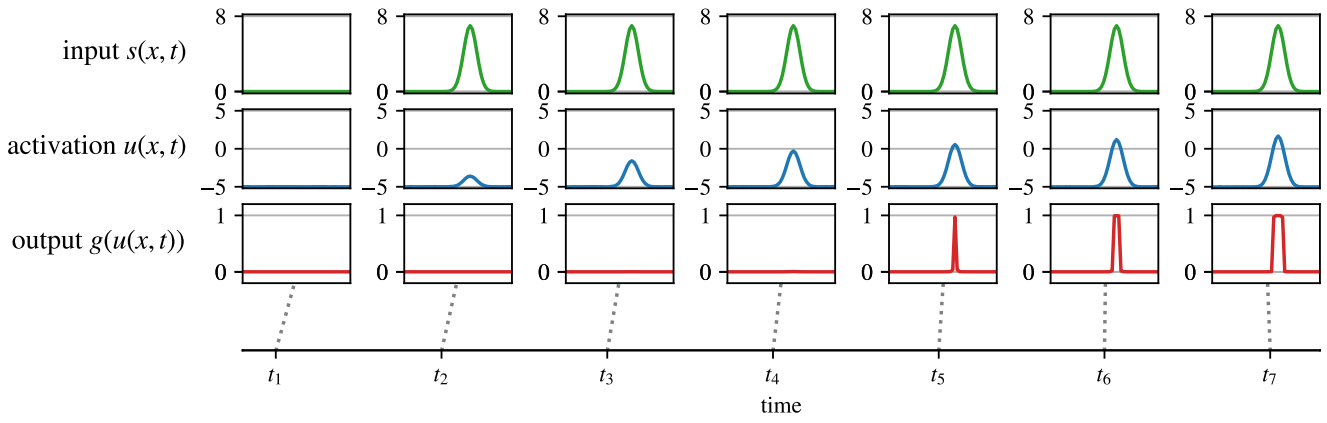


FIGURE 2.15: The detection instability. At time t_2, input is applied, causing the subthreshold attractor to become unstable and an above-threshold attractor to appear. At time t_5, the activation becomes positive, causing the field to yield output. At time t_7, the activation has reached the above-threshold attractor.

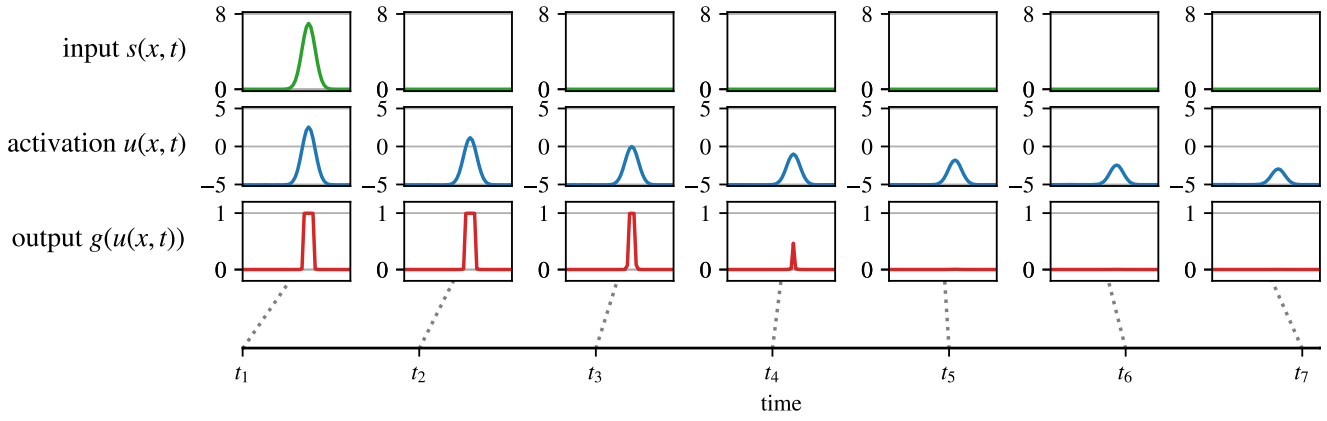


FIGURE 2.16: The reverse detection instability. At time t_2, input has been removed, causing the above-threshold attractor to become unstable and the subthreshold attractor to reappear. At time t_4, the activation has become negative, causing output to cease. It takes time for the field to settle again into the subthreshold attractor.

2.9.4 Instabilities

While fields and nodes evolve in continuous time, qualitative shifts of behavior can occur at discrete moments in time through dynamic instabilities.

2.9.4.1 Detection instability

When localized input of sufficient strength is applied in a certain region, lateral interactions become effective in that region. Positions in that region locally excite each other and inhibit more distant regions. This makes the subthreshold

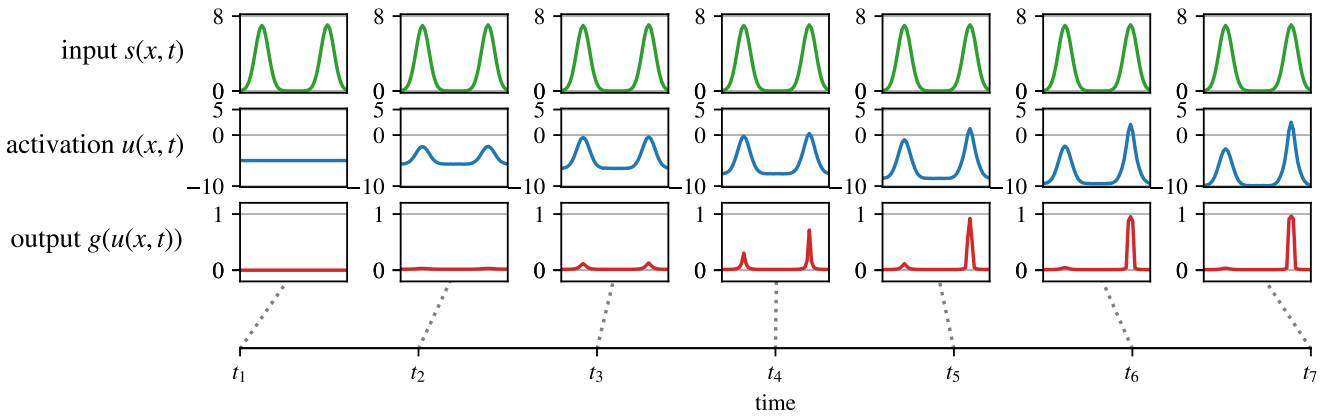


FIGURE 2.17: The selection instability. At time t_1, two sources of input have appeared, causing the field to start to form subthreshold bumps of activity. At time t_4, the activation of the right bump has become positive, causing the field to yield output in that region. Through global inhibition, the left bump is suppressed, preventing it from becoming a peak of positive activation.

attractor disappear and creates an above-threshold attractor, causing the field to form a peak of positive activation in that region (Figure 2.15). This bifurcation is called the *detection instability*, since a way to look at it is as the detection of a significant source of input. When the strength of localized input goes below a critical threshold again, the field goes through the *reverse detection instability*, in which the above-threshold attractor disappears, causing the peak to dissolve (Figure 2.16).

2.9.4.2 Selection instability

As mentioned previously, a field with strong global inhibition can only form a single peak. When multiple regions simultaneously receive localized input with similar strength, the peak that happens to reach a positive activation level first suppresses all other peaks, causing the field to make a selection decision (Figure 2.17). We refer to this instability as the *selection instability*.

2.9.4.3 Working memory

If the amount of local excitation in a field is strong enough, fields can form *self-sustained peaks*, i.e., peaks that remain stable even after the input in their region is removed (Figure 2.18). In other words, self-sustained peaks are peaks for

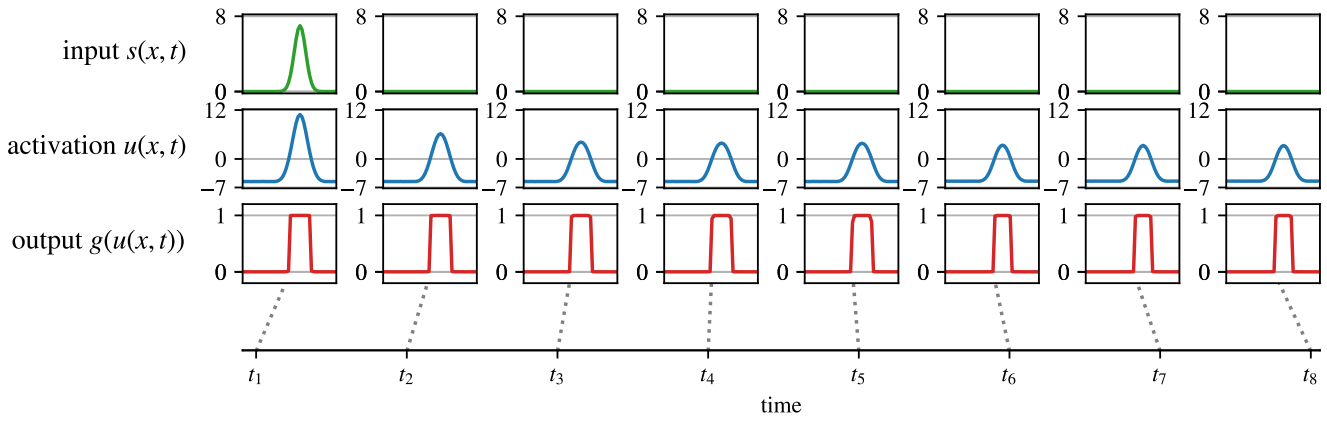


FIGURE 2.18: Working memory. At t_2, input to the field is removed, causing the peak to become weaker. However, due to self-excitation, the peak never goes through the reverse detection instability, but instead remains positive and continues yielding an output.

which the reverse detection instability does not occur. This allows the fields to implement a form of working memory.

2.9.4.4 Instabilities for DNNs

DNNs feature similar dynamic instabilities as DNFs. When input of sufficient strength is applied, a DNN can go through a detection instability and, conversely, through a reverse detection instability upon removal of the input. DNNs can thus have two qualitatively different kinds of states. When they have a positive activation value and consequently an output close to 1, they can be regarded as *active*. When they have a negative activation value and consequently an output close to 0, they can be regarded as *inactive*. With sufficient self-excitation, a DNN can be made *self-sustained*, such that it remains active upon input removal.

DFT postulates that the primitive mechanisms of detection, selection and memory are the basic cognitive mechanisms, and that complex cognitive processes emerge from these three basic mechanisms in coupled architectures.

2.9.5 Coupling

As hinted at already, fields can be coupled with other fields to build complex cognitive architectures. More precisely, the output of a field or node A can serve as input to a field

or node B. Depending on the dimensionalities of A and B, different forms of coupling are possible.

One-to-one — The simplest form of coupling is a *one-to-one coupling*, in which the output of A, multiplied by a weight w, directly serves as input to B according to the equation

$$s_{B,A}(\vec{x}, t) = w \cdot g(u_A(\vec{x}, t)). \tag{2.14}$$

This form of coupling requires that A and B have the same dimensionality.

Expansion — If A has a smaller dimensionality than B, then the vector $\vec{x}_B$ over which B is defined has some additional dimensions to the vector $\vec{x}_A$ over which A is defined. In this case, the entries of the dimensions shared by $\vec{x}_A$ and $\vec{x}_B$ are held constant across all values of the additional dimensions of $\vec{x}_B$. We refer to this operation as an *expansion*. The input to B is then given by the equation

$$s_{B,A}(\vec{x}_B, t) = w \cdot g(u_A(\vec{x}_A, t)). \tag{2.15}$$

In case A is one-dimensional and B is two-dimensional, this creates a ridge input to B for each peak in A. In case A is two-dimensional and B is three-dimensional, it creates a cylindrical input to B for each peak in A.

Contraction — In the converse case where A has larger dimensionality than B, $\vec{x}_A$ has some additional dimensions to $\vec{x}_B$, which need to be dropped. We refer to this operation as a *contraction*. There are two ways in which this can happen. The first way is to take the integral over the additional dimensions,

$$s_{B,A}(\vec{x}_B, t) = w \cdot \int \dots \int dx_{b+1} \dots dx_a g(u_A(\vec{x}_A, t)), \tag{2.16}$$

where a refers to the dimensionality of A and b to the dimensionality of B. Note that we assumed that the additional dimensions of A are given by the last $(a - b)$ entries of $\vec{x}_A$.

The second way is to take the maximum along the additional dimensions,

$$s_{B,A}(\vec{x}_B, t) = w \cdot \max_{x_{b+1},\dots,x_a} g(u_A(\vec{x}_A, t)). \tag{2.17}$$

Point-spread – Long-range projections in the cortex usually undergo a point-spread: A given location in the parameter space of field A projects not just to the same location in field B, but to a range of nearby locations. Point-spread can be modeled by convolving the output of field A with a Gaussian kernel $k_{\mathrm{B,A}}(\vec{x})$ before it serves as input to field B. In that case, the above equations have to be wrapped inside a convolution.

Global excitation, global inhibition — Nodes may be coupled to fields in two ways. In the first way, node A just projects uniformly to every position in field B,

$$s_{B,A}(\vec{x}_B, t) = w \cdot g(u_A(t)). \tag{2.18}$$

If w is positive, this form of coupling is referred to as a *boost* or *global excitation*. Thus, upon activation of node A, all positions in field B receive excitatory input. This may be used as a mechanism for destabilizing the subthreshold attractor.

If w is negative, this form of coupling is referred to as *global inhibition*. Thus, upon activation of node A, all positions in field B receive inhibitory input. This may be used to destabilize an above-threshold attractor.

Patterned connection — The second way in which a node A may be coupled to a field B is by means of a patterned connection, i.e., a connection in which node A projects non-uniformly to different locations in field B. The synaptic connection weights are modeled as a function $W : \mathbb{R}^{dim(\vec{x}_B)} \rightarrow \mathbb{R}$ whose values $W(\vec{x}_B)$ model the degree to which location $\vec{x}_B$ is excited by node A. The input to field B is then given by

$$s_{B,A}(\vec{x}_B, t) = W(\vec{x}_B) \cdot g(u_A(t)). \tag{2.19}$$

As described in detail in Section 2.9.9, patterned connections may serve to model atomic concepts.

2.9.6 Steerable neural mappings

Steerable neural mappings are neural operations that allow to perform coordinate shifts and are supported by empirical evidence (Schneegans & Schöner, 2012). Given two

Schneegans, S. & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics*, *106*(2), 89–109

fields A and B defined over spatial coordinates, where the peaks in A represent the spatial position of objects and B contains a peak at some reference position, the output of a steerable neural mapping yields the positions of the peaks in A relative to the reference position in B. As such, steerable neural mappings can be regarded as performing a translation operation, i.e., as subtracting the reference position from the positions represented in A.

The basis for a steerable neural mapping is a transformation field T with twice the number of dimensions as the spatial fields. T follows the dynamics

$$\begin{aligned}\tau \dot{u}_T(\vec{x}_A, \vec{x}_B, t) = &-u_T(\vec{x}_A, \vec{x}_B, t) + h \\ &+ [k_{\mathrm{T,T}} * g(u_T)](\vec{x}_A, \vec{x}_B, t) \\ &+ [k_{\mathrm{T,A}} * g(u_A)](\vec{x}_A, t) \\ &+ [k_{\mathrm{T,B}} * g(u_B)](\vec{x}_B, t),\end{aligned} \tag{2.20}$$

The first two lines are the generic DNF equation. The third line features an expansion of the output of field A along the additional dimensions $\vec{x}_B$ and the fourth line features an expansion of the output of field B along the additional dimensions $\vec{x}_A$. The resting level h is chosen in such a way that the field only forms peaks where input from A and B overlaps.

The output of field T is used as input to a field C, which is supposed to hold the object positions from field A relative to the reference position from field B. To achieve this, the coupling from T to C is of a special nature: it consists of a diagonal read-out given by

$$s_{\mathrm{C,T}}(\vec{x}, t) = \int d\vec{p} g(u_T(\vec{x} - \vec{p}, \vec{p}, t)). \tag{2.21}$$

For speeding up numerical simulations, steerable neural mappings are often approximated by convolutions of field A with a kernel given by field B.

2.9.7 Behavioral organization

One of the goals of DFT is to build models that evolve autonomously according to the underlying dynamical system. In particular, the system should not depend on a user sending commands or continually providing input. The emergence of discrete cognitive processes, their temporal

Richter, M., Sandamirskaya, Y., & Schöner, G. (2012). A robotic architecture for action selection and behavioral organization inspired by human cognition. In *2012 IEEE/RSJ international conference on intelligent robots and systems* (pp. 2457–2464). New York, NY: IEEE

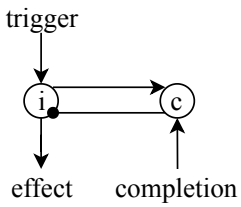


FIGURE 2.19: Elementary behavior. "i": intention node, "c": CoS node

organization, and coordinated use of shared resources need to be accounted for.

DFT architectures are subdivided into *Elementary Behaviors* (EBs), each of which constitutes a functional part of the architecture (Richter et al., 2012). Examples of EBs are motor behaviors (e.g., an arm movement or a saccade), perceptual behaviors (e.g., raising covert attention to an object in the perceptual array) or more abstract cognitive behaviors (e.g., analyzing an attended perceptual object or storing it in memory).

An EB is implemented via two neural nodes (Figure 2.19). An *intention node* represents whether the EB is currently active and able to exert an influence on other parts of the architecture. It may get triggered by external input. For instance, a saccade intention node may get activated upon detection of a salient object in the perceptual array. The connectivity from the intention node to other nodes or fields determines what the effect of this EB will be, e.g., initiating a saccade by boosting a saccade parameter field.

A *Condition of Satisfaction* (CoS) node signals that the EB has successfully finished. It gets activated by external input that signals successful completion of some task. For example, the emergence of a peak in an attention field may signal successful completion of an EB whose goal was to raise attention to an object. Upon activation of the CoS node, it inhibits the intention node, making it inactive. When this happens, the effect that the intention node had on the architecture gets turned off. Sometimes, the fact that an EB has finished needs to be memorized. In this case, the CoS node can be made self-sustained.

The activation u_{IN} of an intention node follows the differential equation

$$\begin{aligned} \tau \dot{u_{\text{IN}}}(t) = &-u_{\text{IN}}(t) + h + s_{\text{IN}}(t) \\ &+ w_{\text{IN,IN}} \cdot g(u_{\text{IN}}(t)) \\ &- w_{\text{IN,CN}} \cdot g(u_{\text{CN}}(t)). \end{aligned} \tag{2.22}$$

The first two lines correspond to the generic DNN equation. $s_{\text{IN}}(t)$ formalizes external input that triggers the elementary behavior, whose source varies with task demands. The third line formalizes the inhibitory input from the CoS node.

The activation of the CoS node follows the differential

equation

$$\begin{aligned}\tau \dot{u_{\text{CN}}}(t) = &-u_{\text{CN}}(t) + h + s_{\text{CN}}(t) \\ &+ w_{\text{CN,CN}} \cdot g(u_{\text{CN}}(t)) \\ &+ w_{\text{CN,IN}} \cdot g(u_{\text{IN}}(t)).\end{aligned} \tag{2.23}$$

Again, the first two lines correspond to the generic DNN equation. $s_{\text{CN}}(t)$ formalizes input that signals the successful completion of the EB, whose source varies with task demands. The third line formalizes input from the intention node.

2.9.8 Serial order

The question of how the nervous system stores serial order of items in memory and triggers serial recall is as of yet unresolved. There are three competing classes of models (Henson & Burgess, 1997). *Interitem association* or *chaining* models assume that sequences are stored via associations between adjacent items, such that the representation of each item in the sequence is associated with the representation of its successor. The activation of an item thus triggers the activation of its successor. *Ordinal* models assume that order is represented along some continuous dimension, e.g., by the relative strengths of the representations, such that the representation of each item in memory is stronger than the representation of its successor in the sequence. The sequence can be retrieved by iteratively selecting the strongest item, and then suppressing it so that it does not get activated again. *Positional* models assume that order is stored by associating each item with a location that encodes its position in the sequence.

Henson, R. & Burgess, N. (1997). Representations of serial order. In *4th neural computation and psychology workshop, london, 9–11 april 1997* (pp. 283–300). Springer

Sandamirskaya and Schöner (2010) argue that empirical evidence supports positional models and propose a neural implementation in the form of a sequence of *ordinal nodes* that get activated one by one (Figure 2.20). Ordinal nodes are self-sustained, such that they remain active until they are actively inhibited. Moreover, the ordinal nodes mutually inhibit each other, resulting in a competition that ensures that only one of them can be active at a time. Each ordinal node excites a self-sustained *memory node*, which serves as a memory of the fact that its associated ordinal node has already been active. Each memory node in turn excites the next ordinal node in the sequence. Due to the

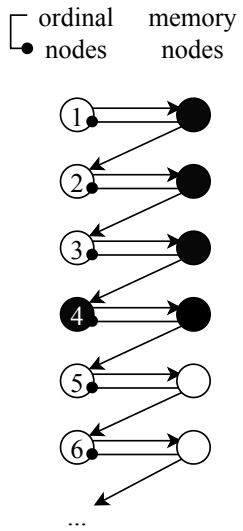


FIGURE 2.20: Serial order mechanism. An exemplary snapshot in time is shown, during which ordinal node 4 is active. Active nodes are black, inactive nodes are white.

Sandamirskaya, Y. & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179

pairwise inhibition between the ordinal nodes, activation of a memory node does not immediately trigger activation of the next ordinal node in the sequence. Instead, the currently active ordinal node has to be actively inhibited in order for the next ordinal node to become active. In order to avoid reactivation of ordinal nodes that have already been active, each memory node slightly inhibits its ordinal node, biasing the competition between the ordinal nodes in favor of the first ordinal node whose predecessor memory node is active but whose own memory node is inactive.

The activation u^i_{ORD} of the ith ordinal node follows the differential equation

$$\begin{aligned}\tau \dot{u}^i_{\text{ORD}}(t) = &-u^i_{\text{ORD}}(t) + h \\ &+ w_{\text{ORD,ORD}} \cdot g(u^i_{\text{ORD}}(t)) \\ &- w_{\text{ORD,ORD}'} \cdot \sum_{i' \neq i} g(u^{i'}_{\text{ORD}}(t)) \\ &+ w_{\text{ORD,MEM}-1} \cdot g(u^{i-1}_{\text{MEM}}(t)) \\ &- w_{\text{ORD,MEM}} \cdot g(u^i_{\text{MEM}}(t)) \\ &- w_{\text{ORD,P}} \cdot g(u_{\text{P}}(t)).\end{aligned} \tag{2.24}$$

The first two lines correspond to the generic DNN equation. The third line formalizes inhibition from other ordinal nodes with strength $w_{\text{ORD,ORD}'}$. The fourth line formalizes excitation by the $i-1$th memory node with strength $w_{\text{ORD,MEM}-1}$. The fifth line formalizes inhibition by the *ith* memory node with strength $w_{\text{ORD,MEM}}$. The last line formalizes inhibition by a *proceed node* with strength $w_{\text{ORD,P}}$, which can be any node whose activation should trigger the next item in a sequence to become active.

The activation u^i_{MEM} of the ith memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}^i_{\text{MEM}}(t) = &-u^i_{\text{MEM}}(t) + h \\ &+ w_{\text{MEM,MEM}} \cdot g(u^i_{\text{MEM}}(t)) \\ &+ w_{\text{MEM,ORD}} \cdot g(u^i_{\text{ORD}}(t)).\end{aligned} \tag{2.25}$$

Again, the first two lines correspond to the generic DNN equation. The third line formalizes excitation of the ith memory node by the ith ordinal node with strength $w_{\text{MEM,ORD}}$.

2.9.9 Concepts

While most theories of concepts describe them at an abstract representational level, previous work by the DFT

research community has provided neural mechanisms for representing, attending to and processing atomic concepts. These mechanisms serve to demonstrate how continuous perceptual representations and discrete conceptual or linguistic representations may be linked.

Attributes are modeled as attribute attention fields. For instance, the attribute COLOR is modeled as a color attention field defined over the continuous, cyclical hue dimension.

Values of an attribute are modeled by DNNs and patterned connections to the attribute field[4]. In line with prototype theories of concepts, the synaptic weight pattern usually corresponds to a Gaussian centered on a prototypical instance of the concept. For example, color concepts are modeled as a DNN (e.g., a red color concept node) that projects to every position in a color attention field. The synaptic weight pattern corresponds to a Gaussian centered on a prototypical hue value for the color category. Upon activation of this node, a peak forms in a region of the color attention field spanning the range of hues corresponding to the color category. Thus, activating a concept node is tantamount to activating or attending to the concept that it represents, which may ultimately lead the cognitive system to attend to objects in the perceptual array falling under that concept.

[4] see Equation 2.19

Conversely, by introducing reverse connections, attending to an object in the perceptual array may activate concept nodes for its attribute values (Richter, 2018); e.g., attending to a red object may activate the red color concept node.

Richter, M. (2018). *A neural dynamic model for the perceptual grounding of spatial and movement relations.* (Doctoral dissertation, Bochum, Ruhr-Universität Bochum)

Classes of relations are modeled by fields. The relations themselves are modeled as DNNs and their synaptic connections with the respective relation field. For instance, the class of spatial relations can be modeled by a spatial relation field, and the spatial relations can be modeled by DNNs (e.g., a below spatial relation concept node) projecting into that field with connectivity weights corresponding to a spatial pattern (see Section 4.3).

The Grounding Strategy Encoder 3

This chapter introduces the *Grounding Strategy Encoder* (GSEnc), a brain system whose job it is to convert combinatorial concepts into a *grounding strategy*, i.e., a sequence of steps that have to be performed in order to ground that concept. This grounding strategy then serves as input to the *Grounding Strategy Executor* (GSEx), which will be introduced in Chapter 4. Its job is to execute the grounding strategy, effectively causing attention to be directed towards an object into the perceptual array that matches the concept. One way to look at this is as supplementing the PA with two further components (Figure 3.1). At the present stage, the GSEnc is only described at a functional level. Future work on the neural basis of transforming conceptual structure into a grounding strategy is needed.

We hypothesize that the grounding of the parts of a combinatorial concept proceeds sequentially as opposed to in parallel. For example, the grounding of Figure 2.3 may proceed by first grounding prototype 4 (i.e., finding a blue object), then grounding prototype 3 (i.e., finding a green diagonal object), then grounding prototype 2 (i.e., finding a red object which is below the object chosen for prototype 3 and above the object chosen for prototype 4), and finally grounding prototype 1 (i.e., finding an object which is to the right of the object chosen for prototype 2). Arguments for this claim are discussed in depth in Section 6.1.1

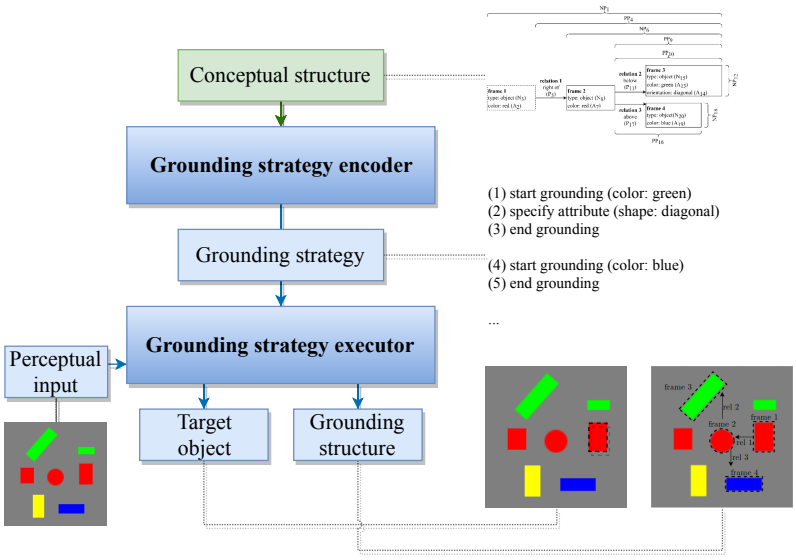


FIGURE 3.1: The language grounding system as an extension of the Parallel Architecture. The conceptual structure feeds into the GSEnc, which outputs a grounding strategy in the form of a sequence of instructions that have to be performed in order to ground the conceptual structure in perception. The grounding strategy feeds into the GSEx that performs the sequence of steps with the result of raising attention to an object in the perceptual array that matches the conceptual structure.

3.1 Target candidate elimination

The process of grounding each frame may be regarded as a form of multiple constraint satisfaction problem. Solving this multiple constraint satisfaction problem may likewise proceed by a sequence of steps that iteratively eliminate target candidates that do not fulfill one of the constraints. After this sequence of steps, only target candidates that fulfill all of the constraints remain, and among those, a selection decision can be made. Arguments for why this is likely to be the way that humans select objects are discussed in Section 6.1.2.

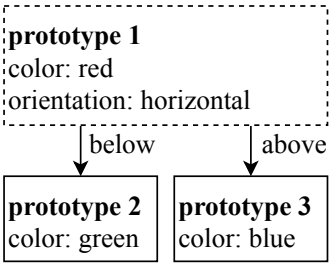


FIGURE 3.2: Exemplary frame graph.

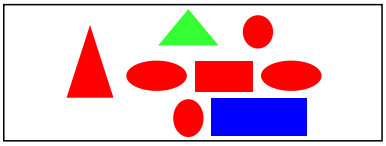


FIGURE 3.3: Exemplary scene.

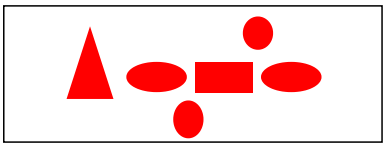


FIGURE 3.4: Target candidates after step 1.

As an example, consider the frame graph expressible by the noun phrase “a red horizontal object below a green object and above a blue object” (Figure 3.2). This frame graph specifies four constraints for the target object to be found: (1) it should be red, (2) it should be horizontal, (3) it should be below a green object, and (4) it should be above a blue object.

The grounding of the red horizontal object in the scene depicted in Figure 3.3 may proceed by first picking a set of red candidate objects (Figure 3.4), then eliminating all candidate objects that are not horizontal (Figure 3.5), then eliminating all candidate objects that are not below the green object (Figure 3.6), and then eliminating all candidate objects that are not above the blue object (Figure 3.7).

```
start grounding (color: green)
specify attribute (orientation: diagonal)
end grounding
start grounding (color: blue)
end grounding
start grounding (color: red)
specify reference (color: green, source: mental
    map)
specify relation (spatial relation: below)
specify reference (color: blue, source: mental
    map)
specify relation (spatial relation: above)
end grounding
start grounding (color: red)
specify reference (color: red, source: mental
    map)
specify relation (spatial relation: right)
end grounding
```

FIGURE 3.8: Exemplary grounding strategy for the combinatorial concept from Figure 2.3 ("a red object right of a red object below a green diagonal object and above a blue object").

3.2 Instruction set

This section describes the types of instructions that may occur in a sequential grounding strategy. Section 6.1.2 discusses theoretical and empirical arguments that motivate this particular choice of the instruction set.

Each instruction in a grounding strategy specifies a task to be performed and, if applicable, an optional set of parameters for the task. While the instructions themselves are abstract elements in a sequential representation, they instruct the GSEx to perform certain tasks. The set of instructions is not agnostic as to how the task is to be performed. Rather, each instruction also determines *how* the grounding system is supposed to perform this task. Thus, each instruction can be characterized by a *purpose*, which describes what the ultimate outcome of the task is, and a *procedure*, which describes how the grounding system is supposed to achieve the purpose. This distinction will become clear when we consider examples.

The possible instructions are summarized in Table 3.1. As a guiding example, we again consider the combinatorial concept describable by the noun phrase "a red object right of a red object below a green diagonal object and above a blue object" (Figure 2.3). Figure 3.8 depicts an exemplary

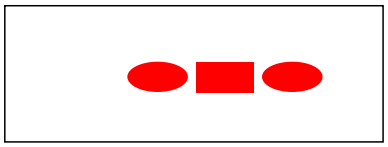


FIGURE 3.5: Target candidates after step 2.

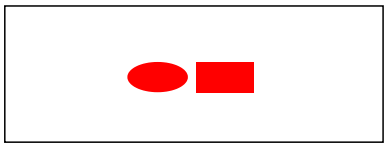


FIGURE 3.6: Target candidates after step 3.

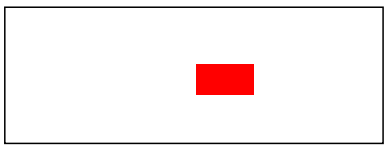


FIGURE 3.7: Target candidates after step 4.

instruction	parameters	purpose	procedure
start grounding	attribute value	a new frame should be grounded; it has the specified attribute value	attend to a set of objects with the specified attribute value and store them in working memory as target candidates
specify attribute	attribute value	the current frame has the specified attribute value	eliminate all target candidates that do not have the specified attribute value
specify reference	attribute value, source	a reference to another object with the specified source and attribute value should be established	attend to an object with the specified attribute value and store it in working memory as a reference object
specify relation	spatial relation	the current frame bears the specified spatial relation to the reference object	eliminate all target candidates that do not bear the specified relation to the reference object
end grounding		the grounding of the current frame is complete	select a target from the remaining target candidates

Table 3.1: Grounding strategy instructions with required parameters, purpose, and procedure.

grounding strategy for that concept.

The start grounding instruction, which is specified in conjunction with an attribute value (a color, an orientation, or a shape), has the purpose of starting the grounding of a new frame with the specified attribute value. It instructs the GSEx to attend to a set of objects with the specified attribute value and to store these objects as target candidate objects in working memory. For example, the instruction `start grounding(color: green)` from Figure 3.8 (line 1) specifies that the grounding of green object should be started, which is achieved by attending to all green objects in the perceptual array and storing them as target candidates. All the instructions after the start grounding instruction and before the next end grounding instruction relate to the same frame, which I shall refer to as the *current frame*.

The specify attribute instruction, which is also specified

in conjunction with an attribute value, has the purpose of declaring that the current frame has the specified attribute value. It instructs the GSEx to eliminate all objects from the set of target candidates in working memory that do not have the specified attribute value. This instruction is repeated for each attribute value in the frame except for the one that was specified in conjunction with the start grounding instruction. For example, the instruction `specify attribute (orientation: diagonal)` from Figure 3.8 (line 2) specifies that the current frame has a diagonal orientation, instructing the GSEx to eliminate all objects that are not diagonal from the green target candidates.

The specify reference instruction, which is specified in conjunction with an attribute value and a source, has the purpose to establish a reference to another object with the specified attribute value and the specified source. This source can be either the scene (i.e., the perceptual input) or the mental map (i.e., a memory store for objects that have been selected as the targets of previously grounded frames). For instance, upon grounding prototype 2 from Figure 2.3, a reference to the object selected for prototype 3 has to be established. Thus, the specify reference instruction instructs the GSEx to attend to an object with the specified attribute value, and to store that object as a reference object in working memory. We hypothesize that even when the source is the mental map, this selection is made on the basis of attribute values as opposed to, say, a pointer to some part of memory. For example, the instruction `specify reference (color: green, source: mental map)` from Figure 3.8 (line 9) specifies that a reference to a green object from the mental map has to be established, which is achieved by attending to a green object from the mental map and storing it as a reference object in working memory. Effectively, this allows the grounding of the current frame (frame 2) to establish a reference to the green diagonal object selected for frame 3. Note that since the selection is made solely based on a single attribute value, the reference object has to be uniquely specified by that attribute value.

The specify relation instruction always follows the specify reference instruction and is specified in conjunction with a relation. Its purpose is to declare that the current frame bears the specified relation to the reference object stored in working memory. It instructs the GSEx to eliminate all

objects from the target candidates in working memory that do not bear the specified relation to the reference object. For instance, the instruction `specify relation (spatial relation: below)` from Figure 3.8 (line 10) specifies that the current frame bears the spatial relation BELOW to the previously specified reference object. This causes the GSEx to eliminate all objects from the red target candidates that are not below the green reference object.

The end grounding instruction has the purpose of declaring that the grounding of the current frame is complete. It instructs the GSEx to select a target object for the current frame among the remaining target candidates in working memory. For instance, Figure 3.8 (line 3) specifies that the grounding of frame 3 is complete, and makes a selection decision among the remaining green diagonal target candidates. Note that the remaining target candidates may contain a single object or no object at all. In the latter case, we say that the grounding process failed.

3.3 Underdetermination of the grounding strategy

A given combinatorial concept does not uniquely determine a grounding strategy. Rather, multiple grounding strategies are possible. This is in line with the introspectively verifiable truth that we have multiple ways to find an object that matches a linguistic description, and that our attention may be guided in different ways.

Firstly, the order in which the frames are grounded may differ. For example, the combinatorial concept from Figure 2.3 may be grounded starting with either the green diagonal object or the blue object. Similarly, any given relation may be processed either forward or backward. For example, the grounding of "a red object below a green object" may proceed either by first attending to a green object and then attending to a red object below it, or by first attending to a red object and then checking if it is below a green object. This fact is also backed up by empirical evidence. While the *Visual World Hypothesis* (VWH) put forward by Tanenhaus et al. (1995) suggests that we attend to objects in the order in which they are mentioned in a sentence, the *Attention Vector Sum Model* (AVS) put forward by

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

Regier and Carlson (2001) predicts that upon grounding a spatial phrase like "a red object below a green object", attention must shift from the green reference object to the red target object. Burigo and Knoeferle (2011) found in an eye-tracking study that participants could flexibly employ both strategies.

The order of instructions within any given frame may also differ. Which of the attributes of a frame is specified with the start grounding instruction as opposed to a specify attribute instruction, and the order in which the attributes of the frame are listed as specify attribute instructions, may be chosen freely without altering the result. In line with the VWH, the attribute mentioned first in a sentence is likely to guide the attribute-based attentional pop-out and is therefore a parameter to the start grounding instruction. The subsequently mentioned attributes are likely to guide candidate elimination in the order in which they are mentioned. For instance, when the noun phrase "a green diagonal rectangle" is processed by a listener, attention is likely to be initially divided between all green objects after hearing the word "green", subsequently narrowed down to all green diagonal objects after hearing the word "diagonal", and finally narrowed down to all green diagonal rectangular objects after hearing the word "rectangular". Alternatively, the choice of the order of attributes may be guided by a heuristic that aims at eliminating as many target candidates in each step as possible. For instance, if there is a large number of green objects in the perceptual input, but only a small number of diagonal objects, then it is more economic to first select all diagonal objects and subsequently eliminate those that are not green. It is likely that the actual order in which candidates are eliminated depends on influences from the order of mention, efficient search heuristics, and a certain degree of randomness.

Regier, T. & Carlson, L. A. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, *130*(2), 273

Burigo, M. & Knoeferle, P. (2011). Visual attention during spatial language comprehension: Is a referential linking hypothesis enough? In L. Carlson, C. H. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society

3.4 A note on implementation

Recall that the GSEnc receives a combinatorial concept as input and transforms it into a sequential grounding strategy. While no neural account of this process is given at the present, it is easy to imagine how this process may proceed algorithmically, namely, by performing a depth-first search on the combinatorial concept graph. Upon visiting

each frame in that search, the attributes and relations in that frame are then merely iterated and transformed into instructions. While this process is easy to implement as a computer program, a neural account for how such a depth-first search or an equivalent procedure may be performed requires careful consideration and thus comprises a future research direction.

The Grounding Strategy Executor 4

This chapter introduces the proposed architecture for the *Grounding Strategy Executor* (GSEx). Given a sequential grounding strategy for a combinatorial concept, the architecture is able to execute that grounding strategy and thus, effectively, to ground the combinatorial concept.

The architecture is depicted in Figure 4.1. It constitutes a single dynamical system, which consists of different components (i.e., nodes and fields), each of which is specified by a differential equation that is coupled to the differential equations of other components. The components of the architecture, along with their activation variables and a summary of the purpose, are listed in Table 4.1. The following sections describe the components in more detail and specify the differential equations that guide their activation.

Due to recurrent interactions among components, each component can only be understood in the context of the other components. However, a description of the components can only be given one component at a time. This is why the description of some of the components may need to refer to components that are introduced in later sections. In order to gain an understanding of the overall architecture, I therefore recommend to continually refer back to Figure 4.1 and Table 4.1. Moreover, when a component is introduced that has been referenced in previous dynamics, I recommend to go back to the previous dynamics. For this purpose, I recommend to use the digital version of this thesis, since it allows to click on the component names to jump to the section where they are defined.

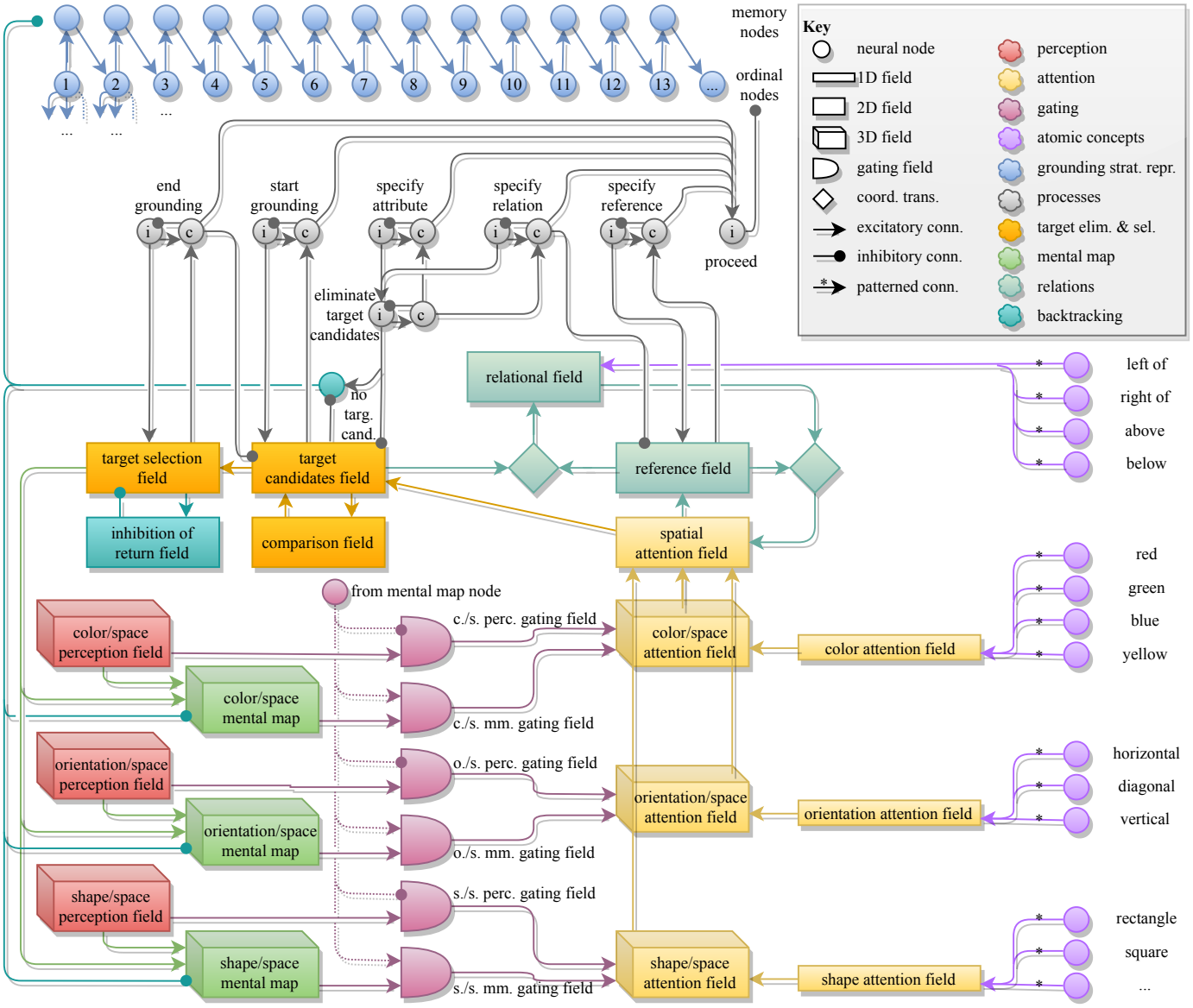


FIGURE 4.1: Overview of the architecture for the grounding of combinatorial concepts. Notational conventions are defined in the upper right.

name	variable	purpose
PERCEPTION		
color/space perception field	$u_{\text{CSPF}}(x, y, c, t)$	object representation over color and space
orientation/space perception field	$u_{\text{OSPF}}(x, y, \phi, t)$	object representation over orientation and space
shape/space perception field	$u_{\text{SSPF}}(x, y, \chi, t)$	object representation over shape and space
ATTENTION		
color attention field	$u_{\text{CAF}}(c, t)$	attention to color values
orientation attention field	$u_{\text{OAF}}(\phi, t)$	attention to orientation values
shape attention field	$u_{\text{SHAF}}(\chi, t)$	attention to shape categories
spatial attention field	$u_{\text{SAF}}(x, y, t)$	attention to spatial locations
color/space attention field	$u_{\text{CSAF}}(x, y, c, t)$	attention to spatial locations and colors
orientation/space attention field	$u_{\text{OSAF}}(x, y, \phi, t)$	attention to spatial locations and orientations
shape/space attention field	$u_{\text{SSAF}}(x, y, \chi, t)$	attention to spatial locations and shapes
GATING		
from mental map node	$u_{\text{FMMN}}(t)$	when inactive, attention is guided to perceptual input; when active, attention is guided to mental map
color/space perception gating field	$u_{\text{CSPGF}}(x, y, c, t)$	gate input from the color/space perception field to the color/space attention field
color/space mental map gating field	$u_{\text{CSMMGF}}(x, y, c, t)$	gate input from the color/space mental map to the color/space attention field
orientation/space perception gating field	$u_{\text{OSPGF}}(x, y, \phi, t)$	gate input from the orientation/space perception field to the orientation/space attention field

orientation/space mental map gating field	$u_{\mathrm{OSMMGF}}(x, y, \phi, t)$	gate input from the orientation/space mental map to the orientation/space attention field
shape/space perception gating field	$u_{\mathrm{SSPGF}}(x, y, \chi, t)$	gate input from the shape/space perception field to the shape/space attention field
shape/space mental map gating field	$u_{\mathrm{SSMMGF}}(x, y, \chi, t)$	gate input from the shape/space mental map to the shape/space attention field
ATOMIC CONCEPTS		
color concept nodes	$u^{\mathrm{R}}_{\mathrm{CCN}}(t)$, $u^{\mathrm{G}}_{\mathrm{CCN}}(t)$, $u^{\mathrm{B}}_{\mathrm{CCN}}(t)$, $u^{\mathrm{Y}}_{\mathrm{CCN}}(t)$	when active, color attention is guided to respective hue
orientation concept nodes	$u^{\mathrm{H}}_{\mathrm{OCN}}(t)$, $u^{\mathrm{D}}_{\mathrm{OCN}}(t)$, $u^{\mathrm{V}}_{\mathrm{OCN}}(t)$	when active, orientation attention is guided to respective angle
shape concept nodes	$u^{\mathrm{R}}_{\mathrm{SCN}}(t)$, $u^{\mathrm{S}}_{\mathrm{SCN}}(t)$, $u^{\mathrm{E}}_{\mathrm{SCN}}(t)$, $u^{\mathrm{C}}_{\mathrm{SCN}}(t)$, $u^{\mathrm{T}}_{\mathrm{SCN}}(t)$	when active, shape attention is guided to respective shape
spatial relation concept nodes	$u^{\mathrm{L}}_{\mathrm{SRCN}}(t)$, $u^{\mathrm{R}}_{\mathrm{SRCN}}(t)$, $u^{\mathrm{A}}_{\mathrm{SRCN}}(t)$, $u^{\mathrm{B}}_{\mathrm{SRCN}}(t)$	when active, a pattern for the respective spatial relation is projected into the spatial relation field
GROUNDING STRATEGY REPRESENTATION		
ordinal nodes	$u^{i}_{\mathrm{ORD}}(t)$	when active, the i-th instruction is processed
memory nodes	$u^{i}_{\mathrm{MEM}}(t)$	when active, the i-th instruction was processed
PROCESSES		
start grounding process intention node	$u_{\mathrm{SGI}}(t)$	when active, the start grounding process will be performed
start grounding process CoS node	$u_{\mathrm{SGC}}(t)$	when active, the start grounding process has just finished
specify attribute process intention node	$u_{\mathrm{SAI}}(t)$	when active, the specify attribute process will be performed
specify attribute process CoS node	$u_{\mathrm{SAC}}(t)$	when active, the specify attribute process has just finished

specify reference process intention node	$u_{\mathrm{SRI}}(t)$	when active, the specify reference process will be performed
specify reference process CoS node	$u_{\mathrm{SRC}}(t)$	when active, the specify reference process has just finished
specify relation process intention node	$u_{\mathrm{SRLI}}(t)$	when active, the specify relation process will be performed
specify relation process CoS node	$u_{\mathrm{SRLC}}(t)$	when active, the specify relation process has just finished
end grounding process intention node	$u_{\mathrm{EGI}}(t)$	when active, the end grounding process will be performed
end grounding process CoS node	$u_{\mathrm{EGC}}(t)$	when active, the end grounding process has just finished
proceed process intention node	$u_{\mathrm{PI}}(t)$	when active, the next ordinal node in the sequence will become active
eliminate target candidates process intention node	$u_{\mathrm{ETCI}}(t)$	when active, target candidates are eliminated
eliminate target candidates process CoS node	$u_{\mathrm{ETCC}}(t)$	when active, target candidates have been eliminated
TARGET CANDIDATE ELIMINATION AND SELECTION		
target candidates field	$u_{\mathrm{TCF}}(x, y, t)$	hold spatial positions of target candidates
comparison field	$u_{\mathrm{CF}}(x, y, t)$	hold peaks at positions of target candidates that are currently attended to
target selection field	$u_{\mathrm{TSF}}(x, y, t)$	select a target from remaining target candidates
MENTAL MAP		
color/space mental map	$u_{\mathrm{CSMM}}(x, y, c, t)$	remember colors and positions of grounded objects
orientation/space mental map	$u_{\mathrm{OSMM}}(x, y, \phi, t)$	remember orientations and positions of grounded objects
shape/space mental map	$u_{\mathrm{SSMM}}(x, y, \chi, t)$	remember shapes and positions of grounded objects
APPREHENDING RELATIONS		

reference field	$u_{\mathrm{RF}}(x, y, t)$	position of a previously grounded reference object
spatial relation field	$u_{\mathrm{SRF}}(x, y, t)$	position of target candidates relative to reference object
BACKTRACKING		
inhibition of return field	$u_{\mathrm{IORF}}(x, y, t)$	remember locations of objects that have been selected as targets
no target candidates node	$u_{\mathrm{NTCN}}(t)$	when active, no target candidates for current frame remain

Table 4.1: Overview of the components (fields and nodes) of the architecture. Each row gives the name, dynamic activation variable (including the dimensions over which the component is defined) and a description of the purpose.

4.1 Perception

The perceptual system receives sensory input from an image or a camera and builds a representation of that input in terms of its features. In doing so, it draws upon the idea of binding through space, according to which the various features of an object are bound together via spatial dimensions that are shared across feature maps (Treisman & Gelade, 1980; Schneegans et al., 2015). Thus, the perceptual system consists of three fields, the color/space perception field, the orientation/space perception field, and the shape/space perception field.

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136

Schneegans, S., Spencer, J., & Schöner, G. (2015). Integrating "what" and "where": Visual working memory for objects in a scene. In G. Schöner & J. Spencer (Eds.), *Dynamic thinking: A primer on dynamic field theory*. New York, NY: Oxford University Press

4.1.1 Color/space perception

The color/space perception field is defined over two spatial dimensions, x, y, and a color dimension c. It can be regarded as capturing the population activation of color-and-space-selective neurons. In our architecture, we remain uncommitted as to the spatial reference frame of that representation (retinal, head-centered, body-centered, or allocentric). Thus, the field can be taken to reflect the activity of early visual cortex populations or parietal lobe populations.

FIGURE 4.2: Exemplary perceptual input.

As a placeholder for a neurally realistic model of visual processing, the input to that field is computed through algorithmic preprocessing of the image: The image is converted to HSV color space, and then converted into a tensor of order 3, defined over spatial location x, y and color c, whose entries scale with the saturation of the objects at the respective locations. This tensor serves as input to the color/space perception field, which is tuned in such a way that sufficiently saturated objects create peaks whose size roughly corresponds to the size of the objects. The activation of those peaks scales with their saturation, capturing bottom-up attention effects.

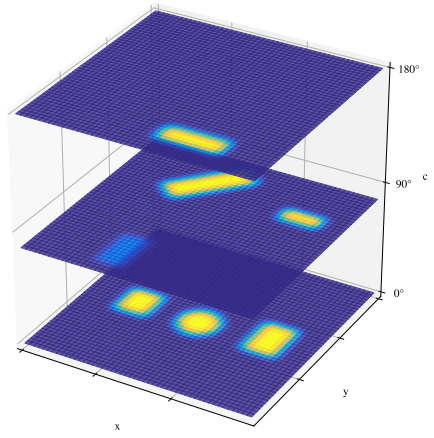


FIGURE 4.3: Activation slices of the color/space perception field in response to the perceptual input from Figure 4.2.

The activation u_{CSPF} of the color/space perception field evolves in time t based on the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{CSPF}}(x, y, c, t) = &- u_{\text{CSPF}}(x, y, c, t) + h \\ &+ [k_{\text{CSPF,CSPF}} * g(u_{\text{CSPF}})](x, y, c, t) \\ &+ [k_{\text{CSPF,C}} * s_{\text{C}}](x, y, c, t).\end{aligned} \tag{4.1}$$

The first two lines correspond to the generic DNF equation. The third line formalizes the input from the algorithmic preprocessing, which is convolved with a kernel $k_{\text{CSPF,C}}$.

Figure 4.3 depicts activation slices of the field for the perceptual input from Figure 4.2.

4.1.2 Orientation/space perception

The orientation/space perception field is defined over two spatial dimensions x, y and an orientation angle dimension ϕ. It can be regarded as capturing the population activation of orientation-and-space-selective neurons, which are to be found, e.g., in visual area V1 or in the parietal lobe, depending on the reference frame.

To compute the input to the orientation/space perception field, the image is converted into HSV color space and the saturation channel is extracted to serve as an intensity profile, formalized as a function $I(x, y)$. The intensity profile is then convolved with filters $F^{\theta}_{\text{Ori}}(x, y)$ for $\theta \in \Theta = \{0°, 45°, 90°, 135°\}$, effectively resulting in a set of intensity distributions for each orientation in Θ.

FIGURE 4.4: Orientation filter $F^{0°}_{\text{Ori}}(x, y)$.

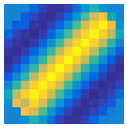


FIGURE 4.5: Orientation filter $F^{45°}_{\text{Ori}}(x, y)$.

FIGURE 4.6: Orientation filter $F^{90°}_{\text{Ori}}(x, y)$.

We model the filters $F^{\theta}_{\text{Ori}}(x, y)$ as a difference of two elliptical Gaussians (Figures 4.4 to 4.6),

$$\begin{aligned} F^{\theta}_{\text{Ori}}(x, y) = a \cdot exp(-(&p_\theta(x - x_0)^2 \\ &+ 2q_\theta(x - x_0)(y - y_0) \\ &+ r_\theta(y - y_0)^2)) \\ &- b \cdot exp(-(p'_\theta(x - x_0)^2 \\ &+ 2q'_\theta(x - x_0)(y - y_0) \\ &+ r'_\theta(y - y_0)^2)), \end{aligned} \tag{4.2}$$

$$p_\theta = \frac{cos^2(\theta)}{2\sigma_x^2} + \frac{sin^2(\theta)}{2\sigma_y^2},$$

$$q_\theta = \frac{sin(2\theta)}{4\sigma_x^2} + \frac{sin(2\theta)}{4\sigma_y^2},$$

$$r_\theta = \frac{sin^2(\theta)}{2\sigma_x^2} + \frac{cos^2(\theta)}{2\sigma_y^2},$$

$$p'_\theta = \frac{cos^2(\theta)}{2\sigma_x'^2} + \frac{sin^2(\theta)}{2\sigma_y'^2},$$

$$q'_\theta = \frac{sin(2\theta)}{4\sigma_x'^2} + \frac{sin(2\theta)}{4\sigma_y'^2},$$

$$r'_\theta = \frac{sin^2(\theta)}{2\sigma'^2_x} + \frac{cos^2(\theta)}{2\sigma'^2_y},$$

$$a > b > 0, \sigma_x < \sigma'_x, \sigma_y < \sigma'_y, \sigma_x > \sigma_y, \sigma'_x > \sigma'_y.$$

These filter profiles exhibit an excitatory center and an inhibitory surround, in line with findings regarding the center-surround receptive-field profile of orientation-selective neurons in the cortex (e.g., the simple cells found in visual area V1; Hubel & Wiesel, 1959).

Hubel, D. H. & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591

Figures 4.7 to 4.9 depict exemplary convolution results for the perceptual input from Figure 4.2. The convolution results $[F^{\phi}_{\text{Ori}} * I](x, y)$ project into the respective regions of the orientation/space perception field, which is tuned in such a way that objects of sufficient intensity create stable peaks of activation whose size roughly corresponds to the size of the object. Formally, the input to the orientation/space perception field is thus given as

$$s_{\text{O}}(x, y, \phi, t) = \begin{cases} [F^{\phi}_{\text{Ori}} * I](x, y) & \text{if } \phi \in \Theta \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

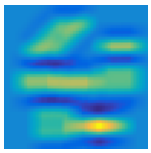


FIGURE 4.7: Convolution result for $\theta = 0°$.

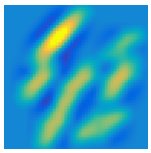


FIGURE 4.8: Convolution result for $\theta = 45°$.

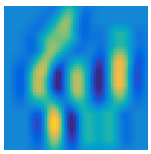


FIGURE 4.9: Convolution result for $\theta = 90°$.

Note that the nervous system cannot directly convolve a perceptual representation with a filter, since this would require synaptic weight sharing among neurons with different receptive fields. Instead, for each filter, multiple neurons responsive to that filter exist, each with a different receptive field. The result of the convolution operation, s_{O}, should just be taken to reflect the population activation of the set of all neurons that are responsive to the filter in any given receptive field.

The activation u_{OSPF} of the orientation/space perception field evolves in time t based on the differential equation

$$\begin{aligned} \tau \dot{u}_{\text{OSPF}}(x, y, \phi, t) = &- u_{\text{OSPF}}(x, y, \phi, t) + h \\ &+ [k_{\text{OSPF,OSPF}} * g(u_{\text{OSPF}})](x, y, \phi, t) \\ &+ [k_{\text{OSPF,O}} * s_{\text{O}}](x, y, \phi, t). \end{aligned} \tag{4.4}$$

The first two lines correspond to the generic DNF equation. The third line formalizes a convolution of the input s_{O} with a kernel k_{OSPF}.

Figure 4.10 depicts activation slices of the field for the perceptual input from Figure 4.2.

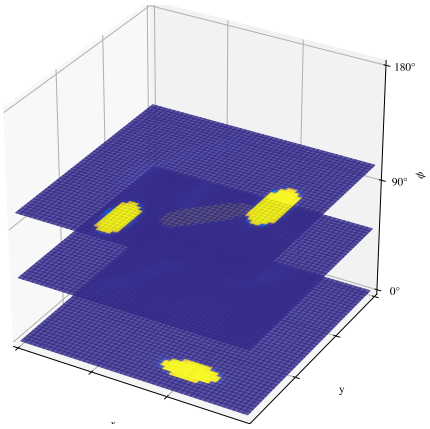


FIGURE 4.10: Activation slices of the orientation/space perception field in response to the perceptual input from Figure 4.2.

4.1.3 Shape/space perception

Nandy, A. S., Sharpee, T. O., Reynolds, J. H., & Mitchell, J. F. (2013). The fine structure of shape tuning in area V4. *Neuron*, *78*(6), 1102–1115

Murata, A., Gallese, V., Luppino, G., Kaseda, M., & Sakata, H. (2000). Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP. *Journal of Neurophysiology*, *83*(5), 2580–2601

The shape/space perception field is defined over two spatial dimensions x, y and a categorical shape dimension χ. It can be regarded as capturing the population activation of shape-and-space-selective neurons, which are to be found, e.g., in visual area V4 (Nandy et al., 2013) or in the parietal lobe (Murata et al., 2000), depending on the reference frame.

Shape patterns are encoded as filters $F^{\chi}_{\text{Shape}}(x, y)$ for $\chi \in= \{\text{R}, \text{S}, \text{E}, \text{C}, \text{T}\}$ denoting, respectively, a rectangle, square, ellipse, circle and triangle filter pattern (see Figures 4.11 to 4.15). These shape patterns feature an excitatory center and an inhibitory surround.

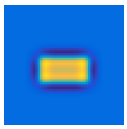


FIGURE 4.11: Rectangle filter.

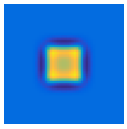


FIGURE 4.12: Square filter.

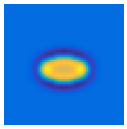


FIGURE 4.13: Ellipse filter.

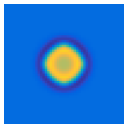


FIGURE 4.14: Circle filter.

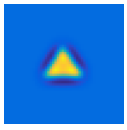


FIGURE 4.15: Triangle filter.

To compute the input to the shape/space perception field, we again take the intensity profile $I(x, y)$ from the saturation channel of the image in HSV color space and convolve it with rotated versions $F^{\chi}_{\text{Shape}}(x, y)$ at angles $\theta \in \Theta = \{0°, 45°, 90°, 135°\}$. The results of all the convolutions for any given shape χ are max-pooled across rotation angles, effectively resulting in a set of intensity distributions for each shape χ that is invariant to rotation. These then project into the respective regions of the shape/space perception field, which is tuned in such a way that objects of sufficient intensity create stable peaks of activation whose size roughly corresponds to the size of the object.

Formally, the input to the shape/space perception field is thus given as

$$\begin{aligned} s_{\text{SH}}(x, y, \chi, t) = \max_{\theta \in \Theta} \int \int g(I(x, y))\, F^{\chi}_{\text{Shape}}(\\ (x - x')cos(\theta) - (y - y')sin(\theta), \\ (x - x')sin(\theta) + (y - y')cos(\theta) \\ )\, dx' dy', \end{aligned} \tag{4.5}$$

which is the maximum filter response across rotations of the filter.

Again, neither the convolution operation nor the rotation operation should be taken to be neural operations. Instead, we assume that for each filter, multiple neurons responsive to that filter exist, each with a different receptive field and rotation angle.

The activation u_{SSPF} of the shape/space perception field

evolves in time t based on the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{SSPF}}(x, y, \chi, t) = &- u_{\text{SSPF}}(x, y, \chi, t) + h \\ &+ \int\int g(u_{\text{SSPF}}(x', y', \chi, t)) \\ &k_{\text{SSPF,SSPF}}(x - x', y - y')dx'dy' \\ &+ \int\int g(s_{\text{SH}}(x', y', \chi, t)) \\ &k_{\text{SSPF,SH}}(x - x', y - y')dx'dy'.\end{aligned} \tag{4.6}$$

The first three lines correspond to the generic DNF equation. Since the shape dimension χ is a categorical dimension, lateral interactions only pertain to the two spatial dimensions x and y. The fourth and fifth line formalize a convolution of the input s_{SH} with a kernel $k_{\text{SSPH,SH}}$, formalizing a point spread that again only pertains to the two spatial dimensions x and y.

Figure 4.17 depicts activation slices of the field for the perceptual input from Figure 4.16.

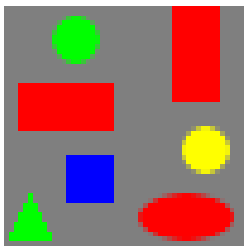


FIGURE 4.16: Exemplary perceptual input.

4.2 Attention

The attentional system consists of seven fields. Three one-dimensional attribute attention fields, the color attention field, the orientation attention field, and the shape attention field, model attention to attribute values. A two-dimensional spatial attention field models attention to spatial locations. The attribute attention fields and the spatial attention field are coupled via three-dimensional attribute/space attention fields, the color/space attention field, the orientation/space attention field and the shape/space attention field.

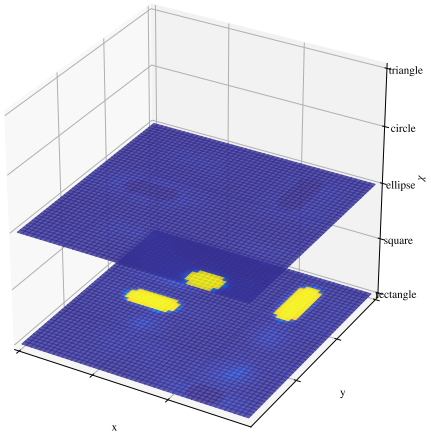


FIGURE 4.17: Activation slices of the shape/space perception field in response to the perceptual input from Figure 4.16.

4.2.1 Attribute attention

Attribute attention fields model attention to metric attribute values like colors or orientations. A peak in such a field reflects the fact that the respective attribute value or range of attribute values is presently being attended to.

The color attention field is defined over the cyclical hue dimension $c \in [0, 360°)$. Its activation u_{CAF} follows the

differential equation

$$\begin{aligned}\tau \dot{u}_{\text{CAF}}(c,t) = &- u_{\text{CAF}}(c,t) + h \\ &+ [k_{\text{CAF,CAF}} * g(u_{\text{CAF}})](c,t) \\ &+ \sum_{C \in \{\text{R,G,B,Y}\}} W^{C}_{\text{Col}}(c) \cdot g(u^{C}_{\text{CCN}}(t))\end{aligned} \tag{4.7}$$

The first two lines correspond to the generic DNF equation. The third line formalizes input from a set of color concept nodes[5] representing discrete color concepts, whose meaning is encoded by the synaptic connection weights W^{C}_{Col}[6]. Upon activation of these nodes, peaks form in the appropriate region of the color attention field, bringing the colors represented by those nodes into attentional focus.

[5] see Section 4.3.1

[6] see Equation 4.22

The orientation attention field is defined over the cyclical orientation angle dimension $\phi \in [0, 360°)$. Its activation follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{OAF}}(\phi,t) = &- u_{\text{OAF}}(\phi,t) + h \\ &+ [k_{\text{OAF,OAF}} * g(u_{\text{OAF}})](\phi,t) \\ &+ \sum_{O \in \{\text{H, D, V}\}} W^{O}_{\text{Ori}}(\phi) \cdot g(u^{O}_{\text{OCN}}(t))\end{aligned} \tag{4.8}$$

The first two lines correspond to the generic DNF equation. Analogously as before, the third line formalizes input from a set of orientation nodes[7] representing discrete orientation concepts, whose meaning is encoded by the synaptic connection weights W^{O}_{Ori}[8]. Again, upon activation of these nodes, peaks form in the appropriate region of the orientation attention field, bringing the orientations represented by those nodes into attentional focus.

[7] see Section 4.3.2

[8] see Equation 4.24

Since shape is a categorical attribute dimension, the shape attention field is defined over the discrete shape dimension $\chi \in \{R, S, E, C, T\}$ denoting, respectively, a rectangle, square, ellipse, circle or triangle shape. Its activation u_{SHAF} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{SHAF}}(\chi,t) = &- u_{\text{SHAF}}(\chi,t) + h \\ &+ w_{\text{SHAF,SHAF}} g(u_{\text{SHAF}}(\chi,t)) \\ &+ w_{\text{SHAF,SCN}} \cdot g(u^{\chi}_{\text{SCN}}(t))\end{aligned} \tag{4.9}$$

The first two lines correspond to the generic DNF equation. Since shape is a discrete dimension, there are no lateral interactions between different shape values. Instead, there is

self-excitation of each shape value with strength $w_{\text{SHAF,SHAF}}$. The third line formalizes input from the respective shape concept node[9].

[9] see Section 4.3.3

4.2.2 Spatial attention

The spatial attention field is defined over the two spatial dimensions x and y. A peak in this field reflects the fact that a certain spatial location is presently being attended to. Its activation u_{SAF} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{SAF}}(x,y,t) = &- u_{\text{SAF}}(x,y,t) + h \\ &+ [k_{\text{SAF,SAF}} * g(u_{\text{SAF}})](x,y,t) \\ &+ \max_{c} ([k_{\text{SAF,CSAF}} * g(u_{\text{CSAF}})](x,y,c,t)) \\ &+ \max_{\phi} ([k_{\text{SAF,OSAF}} * g(u_{\text{OSAF}})](x,y,\phi,t)) \\ &+ \max_{\chi} ([k_{\text{SAF,SSAF}} * g(u_{\text{SSAF}})](x,y,\chi,t)) \\ &+ s_{\text{SAF,SRF}}(x,y,t).\end{aligned} \tag{4.10}$$

The first two lines correspond to the generic DNF equation. The third line formalizes input from the color/space attention field[10], which is contracted along the hue dimension, c. The fourth line formalizes input from the orientation/space attention field, which is contracted along the orientation dimension, ϕ. The fifth line formalizes input from the shape/space attention field, which is contracted along the shape dimension, χ. The last line formalizes input from the spatial relation field, which is transformed into an allocentric coordinate system as described in Section 4.9.

[10] see Section 4.2.3

4.2.3 Attribute/space attention

The attribute/space attention fields are defined over the two spatial dimensions x and y and an additional attribute dimension – color c in the case of the color/space attention field, orientation ϕ in the case of the orientation/space attention field, and shape category χ in the case of the shape/space attention field. Peaks in these fields reflect the fact that an object with a certain attribute value at a certain spatial location is presently being attended to. This way, they serve to couple the attribute attention fields with the spatial attention field.

[11] see below

[12] see Section 4.8

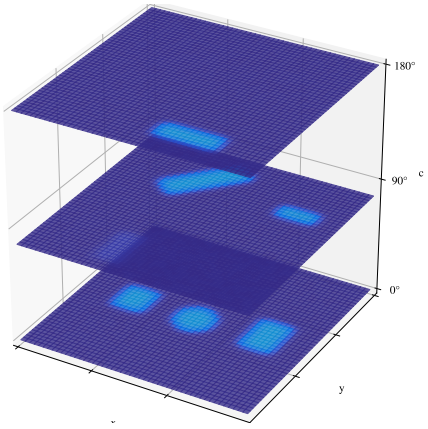


FIGURE 4.18: Activation slices of the color/space attention field in response to the perceptual input from Figure 4.2.

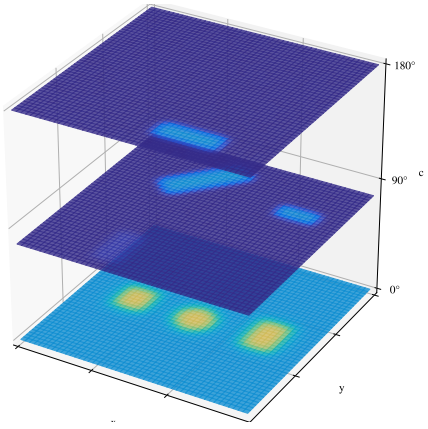


FIGURE 4.19: Activation slices of the color/space attention field in response to the perceptual input from Figure 4.2 when the red color concept node is active.

4.2.3.1 Color/space attention

The color/space attention field is defined over the same dimensions as the color/space perception field, namely, spatial location x, y and color c. Peaks in that field reflect the fact that an object of a certain color at a certain spatial location is currently being attended to. It receives input from the color attention field via an expansion coupling along the shared hue dimension, resulting in a subthreshold ridge of activity when a color is attended to.

Depending on the state of activation of a from mental map node[11], it receives additional input from either the color/space perception field (reflecting objects from the perceptual array) or the color/space mental map[12] (reflecting previously grounded objects stored in memory). This input is too weak to form peaks in the field and thus results in subthreshold bumps of activity (Figure 4.18). When these subthreshold bumps coincide with ridge input from the color attention field, the color/space attention field forms peaks at the appropriate spatial locations and color (Figure 4.19). This way, by attending to a color, objects of that color can be brought into the attentional foreground. These objects can be from the perceptual array (when the from mental map node is inactive) or from the mental map (when the from mental map node is active).

The activation u_{FMMN} of the from mental map node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{FMMN}}(t) = &- u_{\text{FMMN}}(t) + h + s_{\text{FMMN}}(t) \\ &+ w_{\text{FMMN,FMMN}} \cdot g(u_{\text{FMMN}}(t)),\end{aligned} \tag{4.11}$$

which is the generic DNN equation.

To model the fact that the color/space attention field may receive input from either the color/space perception field or the color/space mental map depending on the activation of the from mental map node, two gating fields are introduced. The color/space perception gating field yields output that roughly corresponds to the output of the color/space perception field whenever the from mental map node is inactive, and no output otherwise. Conversely, the color/space mental map gating field yields output that roughly corresponds to the output of the color/space mental map whenever the from mental map node is active, and no output otherwise.

The activation u_{CSPGF} of the color/space perception gat-

ing field follows the differential equation

$$\begin{aligned}\tau\dot{u}_{\text{CSPGF}}(x,y,c,t) = &- u_{\text{CSPGF}}(x,y,c,t) + h \\ &+ [k_{\text{CSPGF,CSPGF}} * g(u_{\text{CSPGF}})](x,y,c,t) \\ &+ [k_{\text{CSPGF,CSPF}} * g(u_{\text{CSPF}})](x,y,c,t) \\ &- w_{\text{CSPGF,FMMN}} \cdot g(u_{\text{FMMN}}(t)).\end{aligned} \tag{4.12}$$

The first two lines correspond to the generic DNF equation. The third line formalizes input from the color/space perception field. The fourth line formalizes global inhibitory input from the from mental map node. The parameters are tuned such that when the from mental map node is active, the gating field yields no output, whereas when the from mental map node is inactive, the output of the gating field roughly corresponds to the output from the color/space perception field.

The activation u_{CSMMGF} of the color/space mental map gating field follows the differential equation

$$\begin{aligned}\tau\dot{u}_{\text{CSMMGF}}(x,y,c,t) = &- u_{\text{CSMMGF}}(x,y,c,t) + h \\ &+ [k_{\text{CSMMGF,CSMMGF}} * g(u_{\text{CSMMGF}})](x,y,c,t) \\ &+ [k_{\text{CSMMGF,CSMM}} * g(u_{\text{CSMM}})](x,y,c,t) \\ &+ w_{\text{CSMMGF,FMMN}} \cdot g(u_{\text{FMMN}}(t)).\end{aligned} \tag{4.13}$$

The first two lines correspond to the generic DNF equation. The third line formalizes input from the color/space mental map[13]. The fourth line formalizes global excitatory input from the from mental map node. The parameters are tuned such that when the from mental map node is inactive, the gating field yields no output, whereas when the from mental map node is active, the output of the gating field roughly corresponds to the output from the color/space mental map.

[13] see Section 4.8

Finally, the activation u_{CSAF} of the color/space attention field follows the differential equation

$$\begin{aligned}\tau\dot{u}_{\text{CSAF}}(x,y,c,t) = &- u_{\text{CSAF}}(x,y,c,t) + h \\ &+ [k_{\text{CSAF,CSAF}} * g(u_{\text{CSAF}})](x,y,c,t) \\ &+ [k_{\text{CSAF,CSPGF}} * g(u_{\text{CSPGF}})](x,y,c,t) \\ &+ [k_{\text{CSAF,CSMMGF}} * g(u_{\text{CSMMGF}})](x,y,c,t) \\ &+ [k_{\text{CSAF,CAF}} * g(u_{\text{CAF}})](c,t).\end{aligned} \tag{4.14}$$

The first two lines correspond to the generic DNF equation. The third line formalizes input from the color/space perception gating field. The fourth line formalizes input from the color/space mental map gating field. The last line formalizes input from the color attention field, which undergoes an expansion coupling along the shared hue dimension, c.

4.2.3.2 Orientation/space attention

The orientation/space attention field implements an analogous mechanism for the orientation attribute. It is defined over the same dimensions as the orientation/space perception field, namely, spatial location x, y and orientation ϕ. Peaks in that field reflect the fact that an object of a certain orientation at a certain spatial location is currently being attended to.

Again, two gating fields are introduced. The orientation/space perception gating field yields output that roughly corresponds to the output of the orientation/space perception field whenever the from mental map node is inactive, and no output otherwise. Its activation u_{OSPGF} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{OSPGF}}(x, y, \phi, t) = &- u_{\text{OSPGF}}(x, y, \phi, t) + h \\ &+ [k_{\text{OSPGF,OSPGF}} * g(u_{\text{OSPGF}})](x, y, \phi, t) \\ &+ [k_{\text{OSPGF,OSPF}} * g(u_{\text{OSPF}})](x, y, \phi, t) \\ &- w_{\text{OSPGF,FMMN}} \cdot g(u_{\text{FMMN}}(t)).\end{aligned} \tag{4.15}$$

The first two lines correspond to the generic DNF equation. The third line formalizes input from the orientation/space perception field. The fourth line formalizes global inhibitory input from the from mental map node.

The orientation/space mental map gating field yields output that roughly corresponds to the output of the orientation/space mental map whenever the from mental map node is active, and no output otherwise. Its activation u_{OSMMGF} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{OSMMGF}}(x, y, \phi, t) = &- u_{\text{OSMMGF}}(x, y, \phi, t) + h \\ &+ [k_{\text{OSMMGF,OSMMGF}} * g(u_{\text{OSMMGF}})](x, y, \phi, t) \\ &+ [k_{\text{OSMMGF,OSMM}} * g(u_{\text{OSMM}})](x, y, \phi, t) \\ &+ w_{\text{OSMMGF,FMMN}} \cdot g(u_{\text{FMMN}}(t)).\end{aligned} \tag{4.16}$$

The first two lines correspond to the generic DNF equation. The third line formalizes input from the orientation/space mental map[14]. The fourth line formalizes global excitatory input from the from mental map node.

[14] see Section 4.8

The activation u_{OSAF} of the orientation/space attention field follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{OSAF}}(x, y, \phi, t) = &- u_{\text{OSAF}}(x, y, \phi, t) + h \\ &+ [k_{\text{OSAF,OSAF}} * g(u_{\text{OSAF}})](x, y, \phi, t) \\ &+ [k_{\text{OSAF,OSPGF}} * g(u_{\text{OSPGF}})](x, y, \phi, t) \\ &+ [k_{\text{OSAF,OSMMGF}} * g(u_{\text{OSMMGF}})](x, y, \phi, t) \\ &+ [k_{\text{OSAF,OAF}} * g(u_{\text{OAF}})](\phi, t).\end{aligned} \tag{4.17}$$

The first two lines correspond to the generic DNF equation. The third line formalizes input from the orientation/space perception gating field. The fourth line formalizes input from the orientation/space mental map gating field. The last line formalizes input from the orientation attention field.

4.2.3.3 Shape/space attention

The shape/space attention field is defined over the same dimensions as the shape/space perception field, namely, spatial location x, y and shape category χ. Peaks in that field reflect the fact that an object of a certain shape at a certain spatial location is currently being attended to.

One more time, two gating fields implement a gating mechanism. The shape/space perception gating field yields output that roughly corresponds to the output of the shape/space perception field whenever the from mental map node is inactive, and no output otherwise. Its activation u_{SSPGF} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{SSPGF}}(x, y, \chi, t) = &- u_{\text{SSPGF}}(x, y, \chi, t) + h \\ &+ \int\int g(u_{\text{SSPGF}}(x', y', \chi, t)) \\ &k_{\text{SSPGF,SSPGF}}(x - x', y - y')dx'dy' \\ &+ \int\int g(u_{\text{SSPF}}(x', y', \chi, t)) \\ &k_{\text{SSPGF,SSPF}}(x - x', y - y')dx'dy' \\ &- w_{\text{SSPGF,FMMN}} \cdot g(u_{\text{FMMN}}(t)).\end{aligned} \tag{4.18}$$

The first three lines correspond to the generic DNF equation. The fourth and fifth line formalize input from the shape/space perception field. The sixth line formalizes global inhibitory input from the from mental map node.

The shape/space mental map gating field yields output that roughly corresponds to the output of the shape/space mental map whenever the from mental map node is active, and no output otherwise. Its activation u_{SSMMGF} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{SSMMGF}}(x, y, \chi, t) = & - u_{\text{SSMMGF}}(x, y, \chi, t) + h \\ & + \int\int g(u_{\text{SSMMGF}}(x', y', \chi, t)) \\ & k_{\text{SSMMGF,SSMMGF}}(x - x', y - y')dx'dy' \\ & + \int\int g(u_{\text{SSMM}}(x', y', \chi, t)) \\ & k_{\text{SSMMGF,SSMM}}(x - x', y - y')dx'dy' \\ & + w_{\text{SSMMGF,FMMN}} \cdot g(u_{\text{FMMN}}(t)).\end{aligned} \tag{4.19}$$

The first three lines correspond to the generic DNF equation. The fourth and fifth line formalize input from the shape/space mental map[15]. The sixth line formalizes global excitatory input from the from mental map node.

[15] see Section 4.8

The activation u_{SSAF} of the shape/space attention field follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{SSAF}}(x, y, \chi, t) = & - u_{\text{SSAF}}(x, y, \chi, t) + h \\ & + \int\int g(u_{\text{SSAF}}(x', y', \chi, t)) \\ & k_{\text{SSAF,SSAF}}(x - x', y - y')dx'dy' \\ & + \int\int g(u_{\text{SSPGF}}(x', y', \chi, t)) \\ & k_{\text{SSAF,SSPGF}}(x - x', y - y')dx'dy' \\ & + \int\int g(u_{\text{SSMMGF}}(x', y', \chi, t)) \\ & k_{\text{SSAF,SSMMGF}}(x - x', y - y')dx'dy' \\ & + w_{\text{SSAF,SHAF}} \cdot g(u_{\text{SHAF}}(\chi, t)).\end{aligned} \tag{4.20}$$

The first three lines correspond to the generic DNF equation. The fourth and fifth line formalize input from the shape/space perception gating field. The sixth and seventh

line formalize input from the shape/space mental map gating field. The last line formalizes input from the shape attention field, which undergoes an expansion coupling along the shared shape dimension, χ.

4.3 Atomic concepts

Following the general DFT mechanism for modeling atomic concepts described in Section 2.6.1, atomic concepts are implemented in the form of neural nodes and their patterned connections to attribute fields or relation fields.

4.3.1 Color concepts

Color concepts are modeled as a set of color concept nodes and their synaptic connections with the color attention field. Thus, upon activating these nodes, peaks form in the appropriate region of the color attention field. For example, activating the red color concept node causes a peak to form in a region of the color attention field which spans the range of red hue values.

The activation variables of the color concept nodes are named u^C_{CCN} with an index $C \in \{\text{R, G, B, Y}\}$ denoting, respectively, the red, green, blue and yellow color concept nodes. They are governed by the differential equation

$$\tau \dot{u}^C_{\text{CCN}}(t) = -u^C_{\text{CCN}}(t) + h + s^C_{\text{CCN}}(t) + w_{\text{CCN,CCN}} \cdot g(u^C_{\text{CCN}}(t)), \tag{4.21}$$

which is the generic DNN equation.

The synaptic weight pattern W^C_{Col} between the color concept node C and the color attention field[16] corresponds to a Gaussian centered on a prototypical hue value for the color category C,

[16] see Equation 4.7

$$\begin{aligned} W^C_{\text{Col}}(c) = exp\left(-\frac{(c-\mu^C_{\text{Col}})^2}{2{\sigma^C_{\text{Col}}}^2}\right), \\ \mu^{\text{R}}_{\text{Col}} = 0^\circ, \sigma^{\text{R}}_{\text{Col}} = 5.04, \\ \mu^{\text{G}}_{\text{Col}} = 120^\circ, \sigma^{\text{G}}_{\text{Col}} = 4.8, \\ \mu^{\text{B}}_{\text{Col}} = 240^\circ, \sigma^{\text{B}}_{\text{Col}} = 10.08, \\ \mu^{\text{Y}}_{\text{Col}} = 60^\circ, \sigma^{\text{Y}}_{\text{Col}} = 3. \end{aligned} \tag{4.22}$$

4.3.2 Orientation concepts

Orientation concepts are modeled as a set of orientation concept nodes and their synaptic connections with the orientation attention field. The activation variables of the orientation concept nodes are named u^O_{OCN} with an index $O \in \{\text{H, D, V}\}$ denoting, respectively, the horizontal, diagonal, and vertical orientation concept nodes. They are governed by the differential equation

$$\tau \dot{u}^O_{\text{OCN}}(t) = -u^O_{\text{OCN}}(t) + h + s^O_{\text{OCN}}(t) + w_{\text{OCN,OCN}} \cdot g(u^O_{\text{OCN}}(t)), \tag{4.23}$$

which is the generic DNN equation.

The synaptic weight patterns W^O_{Ori} between the orientation concept nodes and the orientation attention field[17] correspond to Gaussians centered on a prototypical angle,

[17] see Equation 4.8

$$\begin{aligned} W^O_{\text{Ori}}(\phi) = exp\left(-\frac{(\phi - \mu^O_{\text{Ori}})^2}{2{\sigma^O_{\text{Ori}}}^2}\right), \\ \mu^{\text{H}}_{\text{Ori}} = 0°, \sigma^{\text{H}}_{\text{Ori}} = 6.75, \\ \mu^{\text{D}}_{\text{Ori}} = 45°, \sigma^{\text{D}}_{\text{Ori}} = 6.75, \\ \mu^{\text{V}}_{\text{Ori}} = 90°, \sigma^{\text{V}}_{\text{Ori}} = 6.75. \end{aligned} \tag{4.24}$$

4.3.3 Shape concepts

Shape concepts are modeled as a set of shape concept nodes and their synaptic connections with the shape attention field. The activation variables of the shape concept nodes are named u^χ_{SCN} with an index $\chi \in \{\text{R, S, E, C, T}\}$ denoting, respectively, the rectangle, square, ellipse, circle or triangle shape concept nodes. They are governed by the differential equation

$$\tau \dot{u}^\chi_{\text{SCN}}(t) = -u^\chi_{\text{SCN}}(t) + h + s^\chi_{\text{SCN}}(t) + w_{\text{SCN,SCN}} \cdot g(u^\chi_{\text{SCN}}(t)), \tag{4.25}$$

which is the generic DNN equation.

Since shape concept nodes simply excite the shape attention field at the respective value of χ[18], no synaptic weight patterns are necessary.

[18] see Equation 4.9

4.3.4 Spatial relation concepts

Spatial relation concepts are modeled as a set of spatial relation concept nodes and their synaptic connections with

the spatial relation field[19]. The activation variables of these nodes are called u^S_{SRCN} with an index $S \in \{\text{L, R, A, B}\}$ denoting, respectively, the left, right, above and below spatial relation concept nodes. They are governed by the differential equation

$$\begin{aligned}\tau \dot{u}^S_{\text{SRCN}}(t) = &- u^S_{\text{SRCN}}(t) + h + s^S_{\text{SRCN}}(t) \\ &+ w_{\text{SRCN,SRCN}} \cdot g(u^S_{\text{SRCN}}(t)),\end{aligned} \tag{4.26}$$

which is the generic DNN equation.

Consistent with how humans represent spatial relations (Logan & Sadler, 1996), the synaptic weight pattern W^S_{Spat} between the spatial relation concept nodes and the spatial relation field is modeled as a Gaussian centered on the angle of the direction of that relation (Figures 4.20 to 4.23),

$$\begin{aligned}W^S_{\text{Spat}}(x,y) =& a \cdot exp(\\ & -\frac{(arctan2(y,x) - \mu^S_{\text{Spat}})^2}{2{\sigma^S_{\text{Spat}}}^2} \\ & -\frac{(\sqrt{x^2+y^2} - \mu_r)^2}{2\sigma_r^2} \\ &), \\ & \mu^{\text{L}}_{\text{Spat}} = 180°, \\ & \mu^{\text{R}}_{\text{Spat}} = 0°, \\ & \mu^{\text{A}}_{\text{Spat}} = 90°, \\ & \mu^{\text{B}}_{\text{Spat}} = 270°.\end{aligned} \tag{4.27}$$

4.4 Grounding strategy representation

Recall that the GSEnc transforms a to-be-grounded combinatorial concept into a sequence of instructions that have to be performed in order to ground it. The sequence of instructions then serves as input to the GSEx, which has the task of executing it. This picture is reminiscent of message passing between computer algorithms: One algorithm (the GSEnc) calls another algorithm (the GSEx) with a data structure that comprises a list of instructions. However, since the GSEnc and the GSEx are neural systems as opposed to computer algorithms, the representation of the list of instructions and the message passing have to be explained in neural terms.

[19] see Section 4.9

Logan, G. D. & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (Chap. 13). Cambridge, MA: MIT Press

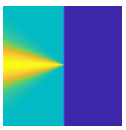


FIGURE 4.20: Spatial pattern $W^{\text{L}}_{\text{Spat}}$.

FIGURE 4.21: Spatial pattern $W^{\text{R}}_{\text{Spat}}$.

FIGURE 4.22: Spatial pattern $W^{\text{A}}_{\text{Spat}}$.

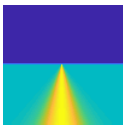


FIGURE 4.23: Spatial pattern $W^{\text{B}}_{\text{Spat}}$.

We assume that the grounding strategy is encoded by means of a positional serial order mechanism as described in Section 2.9.8, i.e., by a set of ordinal nodes that get activated in sequence. Each ordinal node encodes one instruction by virtue of the nodes that it triggers. For instance, the instruction `start grounding (color: green)` is encoded by the fact that its ordinal node is connected to a start grounding process[20] intention node as well as the green color concept node. Thus, upon activation of each ordinal node, an associated process intention node and an optional set of parameter nodes gets activated.

[20] see Section 4.5

As of now, we do not commit to a neural mechanism that establishes the connections between the ordinal nodes and the appropriate process and parameter nodes. Two conceivable alternatives come to mind. First, these connections could correspond to actual synaptic connections. This would require that the GSEnc somehow entrains these connections through fast plasticity. Second, these connections could be mediated by a set of gating neurons $G_{p,i}$, one for each pair of ordinal node i and process intention or parameter node p. The grounding strategy sequence could then be encoded by means of an activity pattern over these nodes, such that $G_{p,i}$ is active if and only if ordinal node i should trigger the activation of node p. $G_{p,i}$ then acts as a gating mechanism for propagating excitation from ordinal node i to node p.

The activation of the ordinal nodes follows Equation 2.24. The activation u^i_{MEM} of the ith memory node follows the equation

$$\begin{aligned}\tau \dot{u}^i_{\text{MEM}}(t) = &-u^i_{\text{MEM}}(t) + h \\ &+ w_{\text{MEM,MEM}} \cdot g(u^i_{\text{MEM}}(t)) \\ &+ w_{\text{MEM,ORD}} \cdot g(u^i_{\text{ORD}}(t)) \\ &- w_{\text{MEM,NTC}} \cdot g(u_{\text{NTC}}(t)).\end{aligned} \tag{4.28}$$

The first 3 lines correspond to the generic memory node equation (Equation 2.25). The last line formalizes inhibitory input from the no target candidates node[21]. As we shall see, this allows the grounding process to go back to the beginning when no target candidates could be found.

[21] see Section 4.10

4.5 Processes

Process organization is implemented based on the principles described in Section 2.9.7. Each process is implemented via two neural nodes: An intention node (labeled "i" in Figure 4.1) represents whether the process is currently active and able to exert an influence on other parts of the architecture. A CoS node (labeled "c") gets activated when the process has successfully finished and inhibits the intention node. In the following, we state the names, functions, and dynamics of these processes. The couplings between these nodes with other components of the architecture, as apparent from the equations of the previous and upcoming sections, make it clear how they actually achieve their function.

4.5.1 Processes for instructions

There is one process for each of the instructions specified in Section 3.2. The activation variables of the intention nodes of these processes are called $u_{P\mathrm{I}}$ with an index

$$P \in \{\mathrm{SG}, \mathrm{SA}, \mathrm{SR}, \mathrm{SRL}, \mathrm{EG}\}$$

denoting, respectively, the start grounding process, specify attribute process, specify reference process, specify relation process and end grounding process. They follow the generic intention node equation (Equation 2.22). The input to these intention nodes is given by

$$s_{P\mathrm{I}}(t) = \sum_i w^i_{P\mathrm{I},\mathrm{ORD}} \cdot g(u^i_{\mathrm{ORD}}(t)), \tag{4.29}$$

where $w^i_{P\mathrm{I},\mathrm{ORD}} > 0$ if and only if the ith instruction in the grounding strategy corresponds to the process P. This formalizes the fact that the ith ordinal node excites a process intention node corresponding to the ith instruction.

The activation variables of the CoS nodes of these processes are called u_{PC}. They follow the generic CoS node equation (Equation 2.23). The inputs to these CoS nodes vary from process to process and are specified in the following. Note that these inputs come from fields that are introduced in upcoming sections. They are given here for reasons of text organization. It might be necessary to refer back to these equations later in order to understand them.

The start grounding process begins the grounding of a new frame. It has to be activated in conjunction with an attribute concept node reflecting an attribute value of the object that is to be grounded. The input to its CoS node is given by

$$s_{\mathrm{SGC}}(t) = w_{\mathrm{SGC,TCF}} \cdot \max_{x,y} g(u_{\mathrm{TCF}}(t)), \tag{4.30}$$

which is a contraction coupling from the target candidates field[22] with activation u_{TCF} multiplied by a synaptic weight $w_{\mathrm{SGC,TCF}}$. Thus, the CoS node of the start grounding process gets activated once a peak has formed in the target candidates field.

[22] see Section 4.6

The specify attribute process allows to specify further attribute values and also has to be activated in conjunction with an attribute concept node. The input to its CoS node is given by

$$s_{\mathrm{SAC}}(t) = w_{\mathrm{SAC,ETCC}} \cdot g(u_{\mathrm{ETCC}}(t)), \tag{4.31}$$

which is excitatory input from the CoS node of the eliminate target candidates process[23]. Thus, the CoS node of the specify attribute process gets activated once the eliminate target candidates process has finished.

[23] see Section 4.5.2

The specify reference process brings an object into the role of the reference object. It is also activated in conjunction with an attribute concept node, which specifies an attribute value of the reference object. The input to its CoS node is given by

$$s_{\mathrm{SRC}}(t) = w_{\mathrm{SRC,RF}} \cdot \max_{x,y} g(u_{\mathrm{RF}}(t)), \tag{4.32}$$

which is a contraction coupling from the reference field[24] with activation u_{RF} multiplied by a synaptic weight $w_{\mathrm{SRC,RF}}$. Thus, the CoS node of the start grounding process gets activated once a peak has formed in the reference field.

[24] see Section 4.9

The specify reference process has to be followed in the sequence by the specify relation process, which is activated in conjunction with a spatial relation that the current frame bears to the reference object. The input to its CoS node is given by

$$s_{\mathrm{SRLC}}(t) = w_{\mathrm{SRLC,ETCC}} \cdot g(u_{\mathrm{ETCC}}(t)), \tag{4.33}$$

which is excitatory input from the CoS node of the eliminate target candidates process[25]. Thus, the CoS node of the

[25] see Section 4.5.2

specify relation process gets activated once the eliminate target candidates process has finished.

The end grounding process signals that the grounding of the current frame is complete and evokes a selection decision. The input to its CoS node is given by

$$s_{\mathrm{EGC}}(t) = w_{\mathrm{EGC,TSF}} \cdot \max_{x,y} g(u_{\mathrm{TSF}}(t)), \tag{4.34}$$

which is a contraction coupling from the target selection field[26] with activation u_{TSF} multiplied by a synaptic weight $w_{\mathrm{EGC,TSF}}$. Thus, the CoS node of the end grounding process gets activated once a peak has formed in the target selection field.

[26] see Section 4.7

4.5.2 Further processes

The proceed process does not correspond to any instruction in the sequence. Instead, its purpose is to cause the next ordinal node in the sequence to become active, which it achieves by inhibiting all ordinal nodes[27]. The activation u_{PI} of its intention node follows the generic intention node equation (Equation 2.22). The input to that node is given by

[27] see Equation 2.24

$$\begin{aligned} s_{\mathrm{PI}}(t) = & w_{\mathrm{PI,SGC}} \cdot g(u_{\mathrm{SGC}}(t)) \\ & + w_{\mathrm{PI,SAC}} \cdot g(u_{\mathrm{SAC}}(t)) \\ & + w_{\mathrm{PI,SRC}} \cdot g(u_{\mathrm{SRC}}(t)) \\ & + w_{\mathrm{PI,SRLC}} \cdot g(u_{\mathrm{SRLC}}(t)) \\ & + w_{\mathrm{PI,EGC}} \cdot g(u_{\mathrm{EGC}}(t)). \end{aligned} \tag{4.35}$$

Thus, it is excited by the CoS nodes of the processes for the instructions. This causes the successful completion of any of these processes to trigger the activation of the intention node of the proceed process, which in turn results in activating the next ordinal node due to inhibition of the current ordinal node. Since this effectively causes the old process intention node to become inactive, its CoS also becomes inactive, and thus the intention node of the proceed process becomes inactive as well. This is why the proceed process does not need a CoS node.

The eliminate target candidates process eliminates target candidates that do not have a specified attribute value or that do not bear a specified relation to a reference object[28]. The activation u_{ETCI} of its intention node follows the generic

[28] see Section 4.6

intention node equation (Equation 2.22). The input to that node is given by

$$\begin{aligned} s_{\text{ETCI}}(t) =& w_{\text{ETCI,SAI}} \cdot g(u_{\text{SAI}}(t)) \\ &+ w_{\text{ETCI,SRLI}} \cdot g(u_{\text{SRLI}}(t)), \end{aligned} \tag{4.36}$$

where the first line is excitatory input from the intention node of the specify attribute process and the second line is excitatory input from the intention node of the specify relation process. Thus, the eliminate target candidates process gets triggered whenever one of these processes is activated. The CoS node of the eliminate target candidates process with activation u_{ETCC} receives no external input, but has a relatively high time constant τ, causing it to become active after a fixed time. As formalized in Equation 4.31 and Equation 4.33, it excites the CoS nodes of the specify attribute process and the specify relation process. Hence, the successful completion of these processes depends on completing the elimination of target candidates.

4.6 Target candidates

The target candidates field has the purpose of holding peaks at the spatial locations of all objects which, at the current stage of processing, are still viable candidates for the frame that is currently poised to be grounded.

Upon starting the grounding of a new frame with the start grounding instruction, it is filled with an initial set of target candidates determined by the attribute concept that is activated in conjunction with that instruction. Subsequently, with each new specify attribute instruction or specify relation instruction, these target candidates are eliminated until only target candidates consistent with all specified attribute values and relations remain.

For example, upon grounding prototype 1 from Figure 2.3 with the sequence of grounding instructions from Figure 3.8 (lines 1-3), the target candidates field should initially be empty. After processing line 1, it should hold peaks at the spatial location of all green objects. Upon processing line 2, all peaks that do not correspond to the spatial locations of diagonal objects should decay, thus leaving only the spatial locations of all green diagonal objects. Similarly, upon grounding prototype 2 from Figure 2.3 with the sequence of grounding instructions from Figure 3.8

(lines 8-13), the target candidates field should again initially be empty. After processing line 8, it should hold peaks at the spatial location of all red objects. After processing lines 9 and 10, all peaks corresponding to objects that are not below the previously grounded green object should be eliminated. After processing lines 11 and 12, all peaks corresponding to objects that are not above the previously grounded blue object should be eliminated.

When the intention node of the start grounding process is active, homogeneous input brings the target candidates field into a dynamic regime where it can form peaks. Input from the spatial attention field determines where peaks arise. Lateral interactions make the target candidates field self-sustained, such that peaks remain in the absence of external input. When peaks have formed in the target candidates field, the CoS node of the start grounding process gets activated[29].

[29] see Equation 4.30

The activation u_{TCF} of the target candidates field follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{TCF}}(x,y,t) = &- u_{\mathrm{TCF}}(x,y,t) + h \\ &+ [k_{\mathrm{TCF,TCF}} * g(u_{\mathrm{TCF}})](x,y,t) \\ &+ [k_{\mathrm{TCF,SAF}} * g(u_{\mathrm{SAF}})](x,y,t) \\ &+ w_{\mathrm{TCF,SGI}} \cdot g(u_{\mathrm{SGI}}(t)) \\ &- w_{\mathrm{TCF,ETCI}} \cdot g(u_{\mathrm{ETCI}}(t)) \\ &+ [k_{\mathrm{TCF,CF}} * g(u_{\mathrm{CF}})](x,y,t) \\ &- w_{\mathrm{TCF,EGC}} \cdot g(u_{\mathrm{EGC}}(t)).\end{aligned} \tag{4.37}$$

The first two lines correspond to the generic DNF equation. The third line formalizes input from the spatial attention field. The fourth line formalizes global excitatory input from the intention node of the start grounding process. The fifth line formalizes global inhibitory input from the intention node of the eliminate target candidates process. The sixth line formalizes input from the comparison field[30]. Together, the fifth line and the sixth line cause peaks in the target candidates field to decay upon activation of the eliminate target candidates process unless they receive support from the comparison field. The last line formalizes inhibitory input from the CoS node of the end grounding process, which empties the field for future grounding processes.

[30] see below

The comparison field has the purpose of holding peaks at the positions of all target candidates that are presently being attended. It receives input from the spatial attention

field and the target candidates field, causing peaks to form where input from the two fields overlaps. The activation u_{CF} of the comparison field follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{CF}}(x,y,t) = &- u_{\mathrm{CF}}(x,y,t) + h \\ &+ [k_{\mathrm{CF,CF}} * g(u_{\mathrm{CF}})](x,y,t) \\ &+ [k_{\mathrm{CF,SAF}} * g(u_{\mathrm{SAF}})](x,y,t) \\ &+ [k_{\mathrm{CF,TCF}} * g(u_{\mathrm{TCF}})](x,y,t).\end{aligned} \tag{4.38}$$

The first two lines correspond to the generic DNF equation. The third line formalizes input from the spatial attention field. The last line formalizes input from the target candidates field.

Recall that the target candidates field receives global inhibition from the intention node of the eliminate target candidates process (Equation 4.37, line 5). Moreover, it receives local excitation from the comparison field. The parameters are tuned such that activation of the eliminate target candidates process causes peaks in the target candidates field to decay unless they receive support from the comparison field. Effectively, this causes only those target candidates to remain that currently receive spatial attention. In the case of the specify attribute process, they are those target candidates that have the specified attribute value, since the specify attribute process is activated in conjunction with an attribute concept node, which causes attention to be drawn to all objects with that attribute value. In the case of the specify relation process, they are those target candidates that bear the specified relation to the reference object, as will be described in Section 4.9. The overall mechanism allows to iteratively eliminate target candidates by listing attributes or relations to other objects.

4.7 Target selection

The target selection field has the purpose of selecting a single target object from the set of target candidates. Input from the target candidates field creates subthreshold bumps of activation at the positions of the target candidates. When the intention node of the end grounding process gets activated, homogeneous input brings the target selection field into a dynamic regime where it can form a peak. Global inhibition is high, allowing only a single peak of activation

to form. This way, when multiple target candidates remain, a selection decision is enforced.

When a peak has formed in the target selection field, the CoS node of the end grounding process gets activated[31]. When this happens, we consider the currently processed frame of the frame graph as grounded.

[31] see Equation 4.34

The activation u_{TSF} of the target selection field follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{TSF}}(x,y,t) = &- u_{\text{TSF}}(x,y,t) + h \\ &+ [k_{\text{TSF,TSF}} * g(u_{\text{TSF}})](x,y,t) \\ &+ [k_{\text{TSF,TCF}} * g(u_{\text{TCF}})](x,y,t) \\ &+ w_{\text{TSF,EGI}} \cdot g(u_{\text{EGI}})(t) \\ &- [k_{\text{TSF,IORF}} * g(u_{\text{IORF}})](x,y,t).\end{aligned} \tag{4.39}$$

The first two lines correspond to the generic DNF equation. The third line formalizes input from the target candidates field. The fourth line formalizes global excitatory input from the intention node of the end grounding process. The last line formalizes inhibitory input from the inhibition of return field[32].

[32] see Section 4.10

4.8 Mental map

In order to be able to refer to previously grounded objects later, as is necessary for grounding relations between objects, we assume that each grounded object is stored in a mental map. The color/space mental map is defined over two spatial dimensions and one color dimension and is self-sustained. It receives input from the target selection field and the color/space perception field. If input from these two fields overlaps, the mental map forms a peak. This way, the mental map stores a representation of the position and color of each grounded object.

The activation u_{CSMM} of the color/space mental map follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{CSMM}}(x,y,c,t) = &- u_{\text{CSMM}}(x,y,c,t) + h \\ &+ [k_{\text{CSMM,CSMM}} * g(u_{\text{CSMM}})](x,y,c,t) \\ &+ [k_{\text{CSMM,CSPF}} * g(u_{\text{CSPF}})](x,y,c,t) \\ &+ [k_{\text{CSMM,TSF}} * g(u_{\text{TSF}})](x,y,t) \\ &- w_{\text{CSMM,NTCN}} \cdot g(u_{\text{NTCN}}(t)).\end{aligned} \tag{4.40}$$

The first two lines are the generic DNF equation. The third line formalizes input from the color/space perception field. The fourth line formalizes an expansion coupling from the target selection field along the color dimension c. The last line formalizes inhibitory input from the no target candidates node[33].

[33] see Section 4.10

The orientation/space mental map implements an analogous mechanism for the orientation attribute. Its activation u_{OSMM} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{OSMM}}(x, y, \phi, t) = &- u_{\text{OSMM}}(x, y, \phi, t) + h \\ &+ [k_{\text{OSMM,OSMM}} * g(u_{\text{OSMM}})](x, y, \phi, t) \\ &+ [k_{\text{OSMM,OSPF}} * g(u_{\text{OSPF}})](x, y, \phi, t) \\ &+ [k_{\text{OSMM,TSF}} * g(u_{\text{TSF}})](x, y, t) \\ &- w_{\text{OSMM,NTCN}} \cdot g(u_{\text{NTCN}}(t)).\end{aligned} \tag{4.41}$$

The first two lines are the generic DNF equation. The third line formalizes input from the orientation/space perception field. The fourth line formalizes an expansion coupling from the target selection field along the orientation dimension ϕ. The last line formalizes inhibitory input from the no target candidates node.

Finally, the shape/space mental map implements an analogous mechanism for the shape attribute. Its activation u_{SSMM} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{SSMM}}(x, y, \chi, t) = &- u_{\text{SSMM}}(x, y, \chi, t) + h \\ &+ \int\int g(u_{\text{SSMM}}(x', y', \chi, t)) \\ &k_{\text{SSMM,SSMM}}(x - x', y - y')dx'dy' \\ &+ \int\int g(u_{\text{SSPF}}(x', y', \chi, t)) \\ &k_{\text{SSMM,SSPF}}(x - x', y - y')dx'dy' \\ &+ [k_{\text{SSMM,TSF}} * g(u_{\text{TSF}})](x, y, t) \\ &- w_{\text{SSMM,NTCN}} \cdot g(u_{\text{NTCN}}(t)).\end{aligned} \tag{4.42}$$

The first three lines are the generic DNF equation. The fourth and fifth line formalize input from the shape/space perception field. The sixth line formalizes an expansion coupling from the target selection field along the shape dimension χ. The last line formalizes inhibitory input from the no target candidates node.

```
start grounding (color: red)
specify reference (color: green, source: scene)
specify relation (spatial relation: below)
end grounding
```

FIGURE 4.25: Grounding strategy for the frame graph from Figure 4.24 when selecting the reference object from the scene.

```
start grounding (color: green)
end grounding
start grounding (color: red)
specify reference (color: green, source: mental
    map)
specify relation (spatial relation: below)
end grounding
```

FIGURE 4.26: Grounding strategy for the frame graph from Figure 4.24 when selecting the reference object from the mental map.

4.9 Apprehending relations

When grounding a frame (the *target frame*) that is characterized by a relation to another frame (the *reference frame*), all target candidates for the target frame that do not bear that relation to the object selected for the reference frame have to be eliminated. For example, when grounding "a red object below a green object" (Figure 4.24), the frame for the red object is the target frame and the frame for the green object is the reference frame. Apprehending the relation surmounts to eliminating all red target candidates from the target candidates field that are not below the green object.

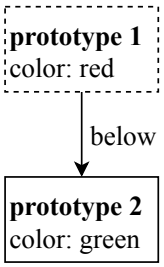


FIGURE 4.24: Frame graph for the query "a red object below a green object".

There are two cases to consider. The first case is to select the reference object directly from the scene (see the grounding strategy in Figure 4.25). This does not allow backtracking[34], and thus should only be done if there is reason to believe that there is only a single reference object with a given attribute value. The second case is that an object for the reference frame has already been selected in a previous grounding step (see the grounding strategy in Figure 4.26). In that case, the reference object is stored in the mental map, and in order to ground the target frame, it is necessary to attend to a previously grounded green object from the mental map, which can be achieved by activating the from mental map node.

The reference field has the purpose of holding a peak at the spatial location of the object selected for the reference frame. It receives global excitatory input from the intention

[34] see Section 4.10

node of the specify reference process, which brings it into a dynamic regime where it can form a peak. Moreover, it receives input from the spatial attention field, causing it to form a peak at the attended spatial location. When this happens, the CoS node of the specify reference process gets activated[35].

[35] see Equation 4.32

Recall from Section 4.5 that the intention node of the specify reference process is always activated in conjunction with an attribute concept node specifying one of the attribute values of the reference object. The spatial attention field thus holds a peak at the spatial location of the reference object, causing the reference field to form a peak at that location.

Recall further from Section 3.2 that the specify reference instruction is always followed by a specify relation instruction. The reference field is therefore self-sustained, such that the position of the reference object is maintained for the next instruction. Upon completion of the specify relation process, global inhibitory input from its CoS node causes the peak in the reference field to decay.

The activation u_{RF} of the reference field follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{RF}}(x,y,t) = &- u_{\mathrm{RF}}(x,y,t) + h \\ &+ [k_{\mathrm{RF,RF}} * g(u_{\mathrm{R}})](x,y,t) \\ &+ [k_{\mathrm{RF,SAF}} * g(u_{\mathrm{SAF}})](x,y,t) \\ &+ w_{\mathrm{RF,SRI}} \cdot g(u_{\mathrm{SRI}}(t)) \\ &- w_{\mathrm{RF,SRLC}} \cdot g(u_{\mathrm{SRLC}}(t)).\end{aligned} \tag{4.43}$$

The first two lines correspond to the generic DNF equation. The third line formalizes input from the spatial attention field. The fourth line formalizes global excitatory input from the intention node of the specify reference process. The last line formalizes global inhibition from the CoS node of the specify relation process.

After a peak has formed in the reference field, the specify relation process has to be activated in conjunction with the concept node of the spatial relation that the target frame bears to the reference frame. The goal now is to attend to the spatial locations of target candidates that bear the specified spatial relation to the reference object.

For this purpose, the representation of the target candidates is transformed into a different coordinate system that is centered on the reference object. This coordinate

transformation can be neurally implemented by a steerable neural mapping (Section 2.9.6); here it is implemented as a convolution to speed up numerical simulations. The output of the transformation is fed into the spatial relation field, which also receives patterned input from the active spatial relation concept node. When these two inputs coincide, peaks can form on the positions of target candidates that bear the specified spatial relation to the reference object. A second coordinate transform converts these peaks back to absolute coordinates and projects into the spatial attention field[36].

[36] see Equation 4.10

The activation u_{SRF} of the spatial relation field follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{SRF}}(x,y,t) = &- u_{\mathrm{SRF}}(x,y,t) + h \\ &+ [k_{\mathrm{SRF,SRF}} * g(u_{\mathrm{SRF}})](x,y,t) \\ &+ \int\int g(u_{\mathrm{TCF}}(x',y',t)) \\ &\cdot g(u_{\mathrm{RF}}(x-x',y-y',t))dx'dy' \\ &+ \sum_{S\in\{L,R,A,B\}} W^S_{\mathrm{Spat}}(x,y) \cdot g(u^S_{\mathrm{SRCN}}(t)).\end{aligned} \tag{4.44}$$

The first two lines correspond to the generic DNF equation. The third and fourth line formalize the approximation of the steerable neural mapping as a convolution of the output of the target candidates field with a kernel given by the output of the reference field. The last line formalizes input from the set of spatial relation concept nodes multiplied by the synaptic connection weights W^S_{Spat} defining the spatial relation patterns[37].

[37] see Section 4.3.4

4.10 Backtracking

Until now, we have assumed that the target selection field can always select a target object for the current frame. However, this presupposes that the target candidates field is not empty. When it is empty, this means that no object could be found which has all the specified attribute values and bears all the specified relations to the reference objects.

This may happen for one of two reasons. The first possible reason is that there is no object in the scene which matches the combinatorial concept. For understanding the

second possible reason, recall from Section 4.7 that when multiple target candidates remain, the target selection field makes an arbitrary selection decision. It may turn out that this selection decision is false. However, this cannot be known at the time when the selection decision is made. It can, however, be inferred based on the fact that in a future grounding process, no target candidates remain, which may indicate that one of the previous grounding steps made a false selection decision, causing a false reference object to be used for target candidate elimination.

In order to identify that no target candidates remain for the current frame, we introduce a no target candidates node, which becomes active whenever target candidate elimination has eliminated all peaks in the target candidates field. It is excited by the intention node of the eliminate target candidates process and inhibited by a contraction coupling from the target candidates field. The parameters are chosen such that the node is active if and only if the eliminate target candidates process is active while there is no peak in the target candidates field, which is the case whenever the eliminate target candidates process has eliminated all target candidates.

The activation u_{NTCN} of the no target candidates node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{NTCN}}(t) = &- u_{\text{NTCN}}(t) + h + w_{\text{NTCN,NTCN}} \cdot g(u_{\text{NTCN}}(t)) \\ &+ w_{\text{NTCN,ETCI}} \cdot g(u_{\text{ETCI}}(t)) \\ &- w_{\text{NTCN,TCF}} \cdot \max_{x,y} g(u_{\text{TCF}}(x, y, t)).\end{aligned} \tag{4.45}$$

The first line corresponds to the generic DNN equation. The second line formalizes excitatory input from the intention node of the eliminate target candidates process. The third line formalizes an inhibitory contraction coupling from the target candidates field.

Recall that the memory nodes receive global inhibitory input from the no target candidates node.[38] Thus, in case no target candidates remain in a grounding process, all memory nodes are deactivated, causing the serial order mechanism to start again from the beginning. This allows the system to make different selection decisions for the frames.

[38] see Equation 4.28

Further recall that the mental map fields receive global inhibitory input from the no target candidates node.[39] This

[39] see Equations 4.40 to 4.42

causes the mental map to be reset completely when the grounding attempt fails, so that the new grounding attempt, which starts again from the beginning of the sequence, can make different selection decisions.

In order to avoid making the same selection decisions in the next attempt at processing the grounding sequence, the inhibition of return field remembers all the target selection decisions. It is defined over the spatial dimensions x and y and is self-sustained. It inhibits the target selection field[40], which biases the competition in favor of objects that have not been tried before. Thus, during the next attempt at processing the grounding sequence, different selection decisions are made.

[40] see Equation 4.39

The activation u_{IORF} of the inhibition of return field follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{IORF}}(x,y,t) = &- u_{\text{IORF}}(x,y,t) + h \\ &+ [k_{\text{IORF,IORF}} * g(u_{\text{IORF}})](x,y,t) \\ &+ [k_{\text{IORF,TSF}} * g(u_{\text{TSF}})](x,y,t).\end{aligned} \tag{4.46}$$

The first two lines correspond to the generic DNF equation. The third line formalizes excitatory input from the target selection field.

Results 5

This chapter presents a number of simulations of the architecture using the software framework *cedar* (Lomp, Zibner, Richter, Ranó, & Schöner, 2013). The various simulations differ only in perceptual input and queries, which can be supplied by the user in the form of the input image to the perceptual fields (see Section 4.1) and the list of instructions and parameters that encode the grounding strategy (see Section 4.5.1). The model parameters are left unchanged across simulations.

Lomp, O., Zibner, S. K. U., Richter, M., Ranó, I., & Schöner, G. (2013). A software framework for cognition, embodiment, dynamics, and autonomy in robotics: Cedar. In *International conference on artificial neural networks* (pp. 475–482). Springer

Queries have been chosen to demonstrate qualitatively different scenarios of varying complexity. They differ in the number of frames, the number of attributes per frame, the number of relations between frames, and the pattern of interdependence between frames. The images have been specifically designed to provide a test bed for the respective queries.

For each query, the time course of activation of nodes and fields is plotted and analyzed qualitatively for its correctness. Criteria for judging the correct behavior include the following:

1. If there is at least one object in the perceptual input that matches the query, then the target selection field contains a peak for this object at the end of the grounding sequence.

2. For each frame that is part of the query, the mental map stores a representation of the object that is matched to that frame.

3. Processes are carried out in the correct order and do not interfere with each other.
4. Processes meet their conditions of satisfaction.
5. Fields form stable representations.
6. At any time, the pattern of activation over nodes and fields coheres with the demands of the currently active process.

Overall, it was found that the model fulfills these criteria in all experiments.

prototype 1
color: red

FIGURE 5.1: Query with a single frame and a single attribute.

FIGURE 5.2: Scene for the query from Figure 5.1.

5.1 Single frame, single attribute

The most simple conceivable query is to search for an object with a single attribute value, e.g., "a red object". This is not yet a combinatorial query, thus the mental map and the relational machinery of the architecture play no role. Nonetheless, it serves as a good introductory example to understand the roles and interactions of some of the fields and processes. Figure 5.1 depicts the trivial frame graph for that query. Figure 5.2 depicts the exemplary scene in which the query is performed. Notice that there are three possible targets for the query. The grounding strategy for that query is given as follows:

```
1 start grounding (color: red)
2 end grounding
```

The time course of activation of relevant nodes and activation snapshots of relevant fields upon performing that query are depicted in Figure 5.3.

At time t_1, ordinal node 1 has caused the red color concept node to become active. As a result, a peak formed in the red region of the color attention field. Through the expansion coupling with the color/space attention field, the red region of that field received homogeneous excitation, which resulted in peaks at the positions and colors of the red objects. Through the contraction coupling with the spatial attention field, peaks formed in that field on the locations of the red objects. Meanwhile, ordinal node 1 has activated the start grounding process intention, which provides global excitation to the target candidates field, bringing it into a dynamic regime where it can form peaks.

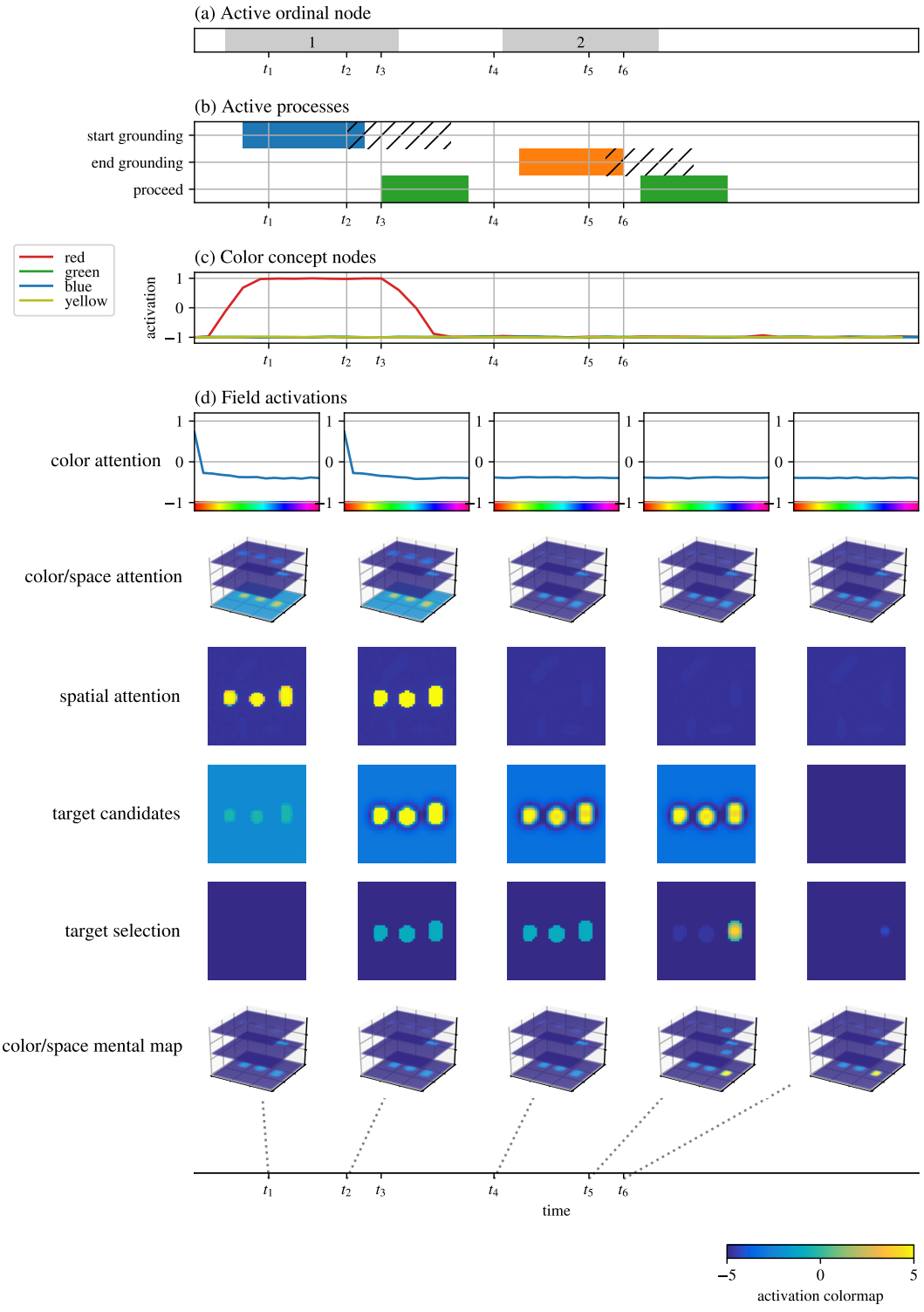


FIGURE 5.3: Time course of relevant parts of the architecture as it grounds the concept in Figure 5.1. The x axes of all figures are time axes and indicate time points of interest (t_1, t_2, ...). **(a)** Active ordinal nodes as rectangles whose range corresponds to the time range of activation. **(b)** Active processes as colored rectangles whose range corresponds to the time range during which the intention node of the respective process is active. The range of the stripe pattern corresponds to the time range during which the CoS node of that process is active. **(c)** Activation time courses of color concept nodes. **(d)** Activation snapshots of relevant fields at time points of interest.

At time t_2, peaks have formed in the target candidates field on the locations of the red objects. As a result, the CoS node of the start grounding process gets activated, which signals that the process has successfully finished. The target candidates have also created subthreshold bumps of activation in the target selection field.

At time t_3, the CoS node of the start grounding process has activated the intention node of the proceed process, which starts inhibiting ordinal node 1.

At time t_4, ordinal node 1 is inactive, and, consequently, the start grounding process and the red color concept node are inactive as well. The target candidates field thus no longer receives global excitation, and the spatial attention field no longer receives input at the locations of the red objects. Nonetheless, due to the self-excitation of the target candidates field, the peaks on the red objects are maintained.

At time t_5, ordinal node 2 has caused the intention node of the end grounding process to become active, which provided global excitation to the target selection field, causing it to form a peak and, thus, to select one of the target candidates as a target object for the query. Through the excitatory expansion coupling with the color/space mental map, it stored an object representation of the selected target object.

At time t_6, the CoS node of the end grounding process has inhibited the target candidates field, thus emptying it for future grounding processes. The color/space mental map has maintained its representation of the selected target object, allowing future grounding processes to refer back to it.

prototype 1
color: red
orientation: horizontal
shape: rectangle

FIGURE 5.4: Query with a single frame and multiple attributes.

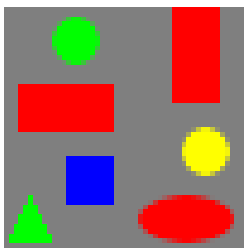


FIGURE 5.5: Scene for the query from Figure 5.4.

5.2 Single frame, multiple attributes

Queries become more interesting when multiple attributes are combined, as in "a red horizontal rectangle". Figure 5.4 depicts the frame graph and Figure 5.5 depicts the scene for that query. The grounding strategy is given as follows:

```
1 start grounding (color: red)
2 specify attribute (orientation: horizontal)
3 specify attribute (shape: rectangle)
4 end grounding
```

This grounding strategy instructs the GSEx to first select a set of red target candidates, then to eliminate all target candidates that are not horizontal, and then to eliminate all target candidates which do not have a rectangle shape.

This query does not yet require the mental map and relational machinery of the architecture, since it contains a single frame that does not refer to any other frames. However, this query serves to demonstrate the sequential target candidate elimination process.

Figure 5.6 depicts the activation time course of relevant nodes and fields upon grounding that frame. At time t_1, ordinal node 1 has activated the red color concept node, which has caused the spatial attention field to form peaks on the red objects. The ordinal node has also activated the start grounding process intention, which has brought the target candidates field into a dynamic regime where it can form peaks, resulting in self-sustained peaks on the positions of the red objects.

At time t_2, ordinal node 2 has activated the horizontal orientation concept node, which has caused the spatial attention field to form peaks on the horizontal objects. As a result, the comparison field has formed peaks on all target candidates that are horizontal. Moreover, ordinal node 2 has activated the specify attribute process intention, which has activated the eliminate target candidates process intention.

At time t_3, the eliminate target candidates process intention has eliminated all target candidates that do not receive support from the comparison field, i.e., all target candidates that are not horizontal. Consequently, the target candidates field retains peaks on all red horizontal objects.

At time t_4, ordinal node 3 has activated the rectangle shape concept node, which has caused the spatial attention field to form peaks on all rectangles. As a result, the comparison field has formed peaks on all target candidates that are rectangles. Moreover, ordinal node 3 has activated the specify attribute process intention, which has activated the eliminate target candidates process intention.

At time t_5, the eliminate target candidates process intention has eliminated all target candidates that do not receive support from the comparison field, i.e., all target candidates that are not rectangles. Consequently, the target candidates field retains peaks on all red horizontal rectangles.

At time t_6, ordinal node 4 has activated the end grounding process intention, causing the target selection field to

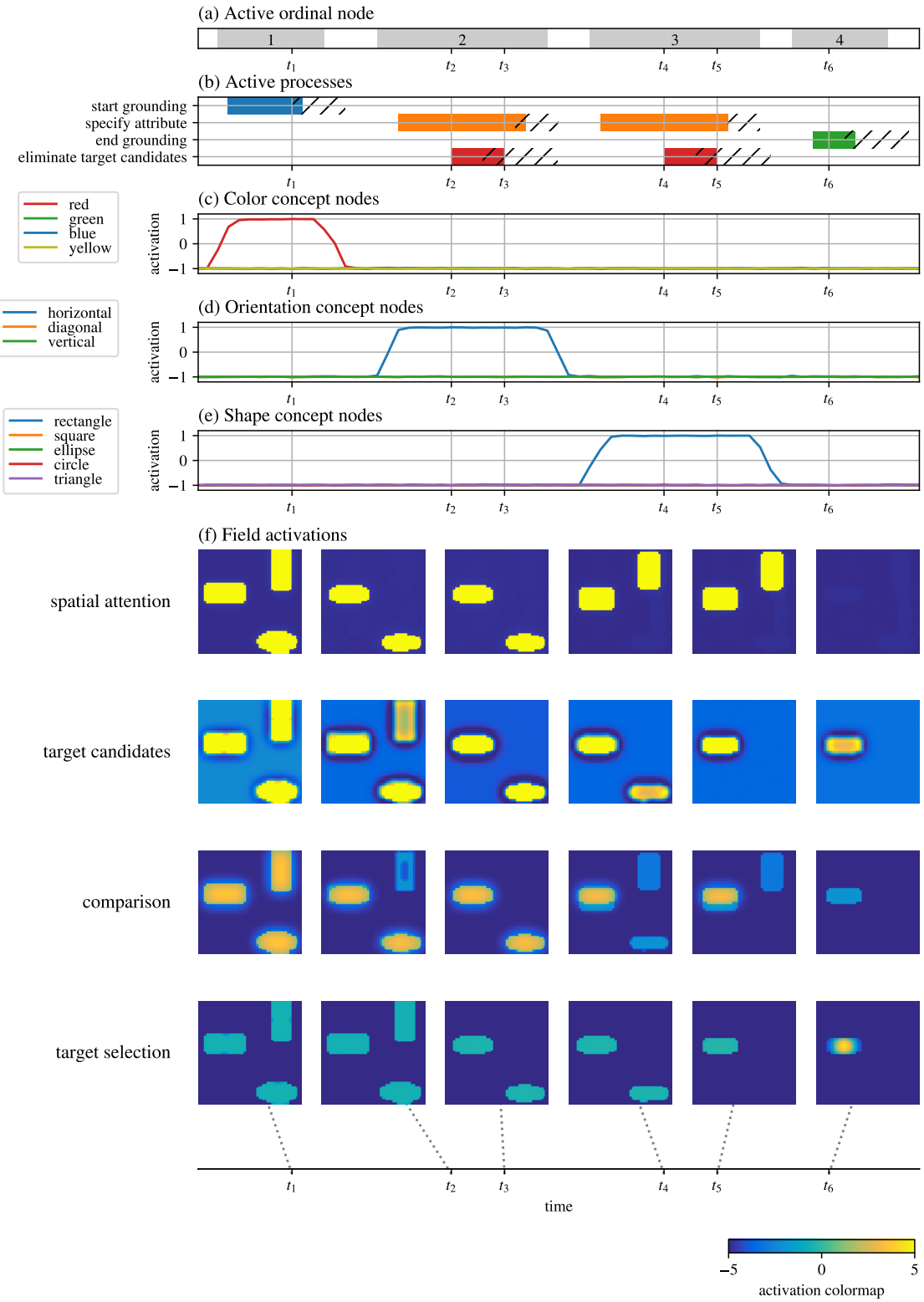


FIGURE 5.6: Time course of relevant parts of the architecture as it grounds the concept in Figure 5.4. **(a)** Active ordinal nodes. **(b)** Active processes. **(c)** Activation time courses of color concept nodes. **(d)** Activation time courses of orientation concept nodes. **(e)** Activation time courses of shape concept nodes. **(f)** Activation snapshots of relevant fields at time points of interest.

form a peak on the red horizontal rectangle.

5.3 Single relation, unambiguous reference

So far, the simulations dealt with single frames without relations. The reference field and the spatial relation field were thus unused. Sometimes, a given object cannot be disambiguated based on attribute values alone. Figure 5.8 depicts a scene where this is the case. There are two red circles. The only way for a linguistic phrase to disambiguate the two is by specifying the spatial relation between the red target object and the unique green reference object, i.e., "a red object below a green object" (Figure 5.7).

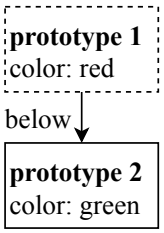


FIGURE 5.7: Query with a single relation.

The grounding strategy is given as follows:

```
1 start grounding (color: red)
2 specify reference (color: green, source: scene)
3 specify relation (spatial relation: below)
4 end grounding
```

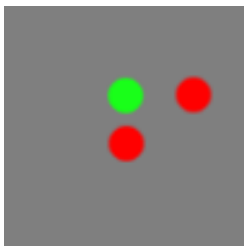


FIGURE 5.8: Scene for the query from Figure 5.7.

It instructs the GSEx to first select a set of red target candidate objects (line 1), then to select a green reference object from the scene and store it in the reference field (line 2), then to guide spatial attention to all target candidates below the reference object (line 3), which in turn causes all target candidates that are not below the reference object to be eliminated, and finally to select the remaining red target candidate in the target selection field (line 4).

Figure 5.9 depicts the activation time course of relevant nodes and fields upon grounding the query. At time t_1, ordinal node 1 has activated the start grounding process intention and the red color concept node, causing the two red target candidates to be stored in the target candidates field.

At time t_2, ordinal node 2 has activated the specify reference process intention, which has brought the reference field into a dynamic regime where it can form a peak. Moreover, it has activated the green color concept node, which has caused spatial attention to be directed to the green object. Consequently, the position of the green object is stored in the reference field.

At time t_3, ordinal node 3 has activated the below spatial relation concept node, which has caused the spatial below

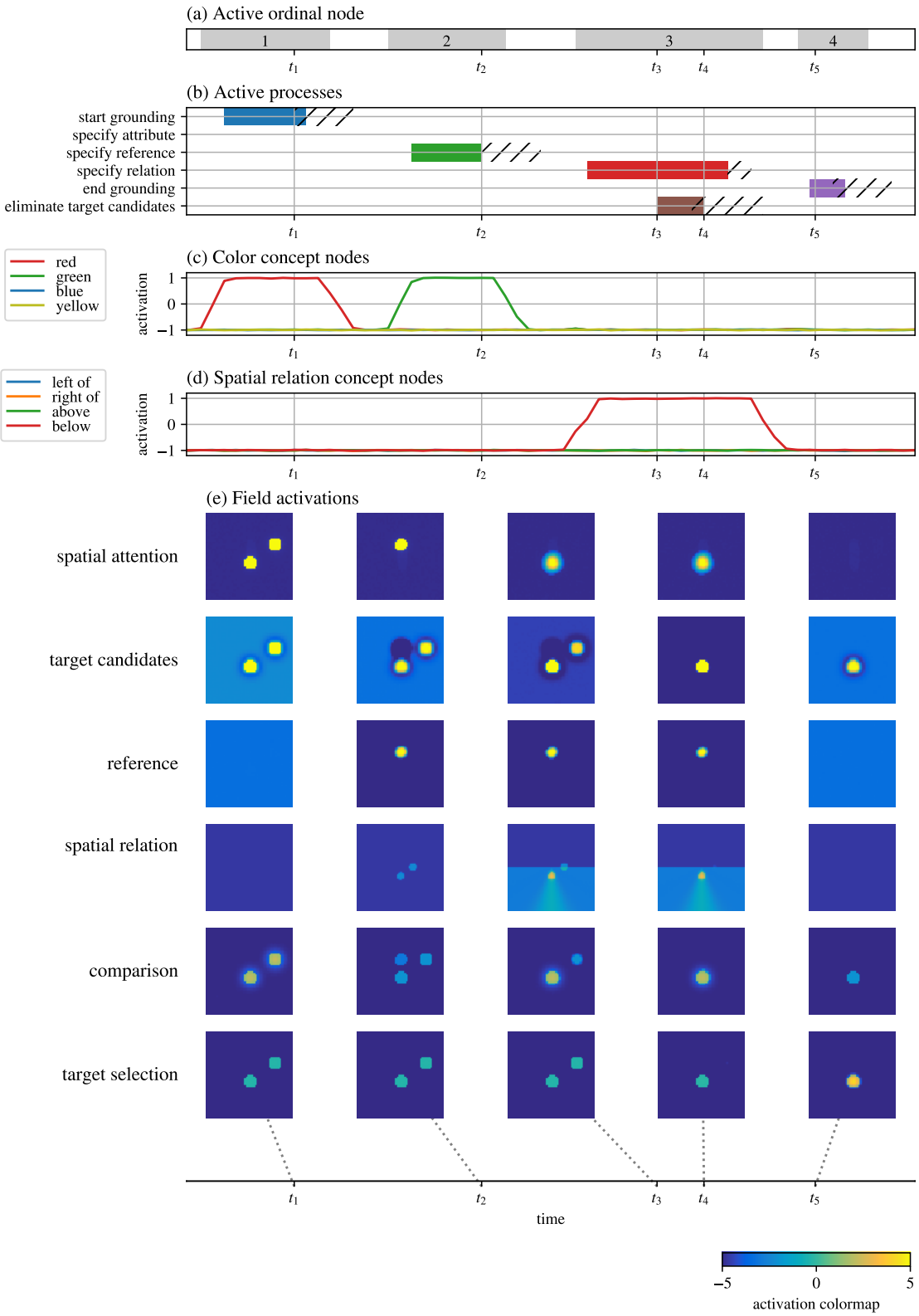


FIGURE 5.9: Time course of relevant parts of the architecture as it grounds the concept in Figure 5.7. **(a)** Active ordinal nodes. **(b)** Active processes. **(c)** Activation time courses of color concept nodes. **(d)** Activation time courses of spatial relation concept nodes. **(e)** Activation snapshots of relevant fields at time points of interest.

pattern to be fed into the spatial relation field. This resulted in a peak on the relative position of the red object below the green object. Through the reverse coordinate transformation, a peak on the absolute position of that object formed in the spatial attention field. As a result, the comparison field formed a peak on that position. Ordinal node 3 has also activated the specify relation process intention, which has in turn activated the eliminate target candidates process intention.

By time t_4, the eliminate target candidates process intention has eliminated all target candidates that did not receive excitatory support from the comparison field, i.e., all target candidates that are not below the green object, leaving the single red object below the green object.

At time t_5, the red object is selected in the target selection field.

5.4 Multiple relations

Assume we want to refer to the red object in the middle of Figure 5.11. This object can neither be disambiguated based on attributes, nor based on a single relation. Instead, it has to be disambiguated by the fact that it is a red object that is both below a green object and above a blue object. Figure 5.10 depicts the frame graph for that query.

The grounding strategy is given as follows:

```
1 start grounding (color: red)
2 specify reference (color: green, source: scene)
3 specify relation (spatial relation: below)
4 specify reference (color: blue, source: scene)
5 specify relation (spatial relation: above)
6 end grounding
```

The activation time courses of relevant nodes and fields upon grounding that query are depicted in Figure 5.12. At time t_1, ordinal node 1 has activated the start grounding process intention and the red color concept node, causing the three red objects to be stored in the target candidates field.

At time t_2, ordinal node 2 has activated the specify reference process intention and the green color concept node, causing the green object to be stored in the reference field.

At time t_3, ordinal node 3 has activated the specify relation process intention and the below spatial relation concept

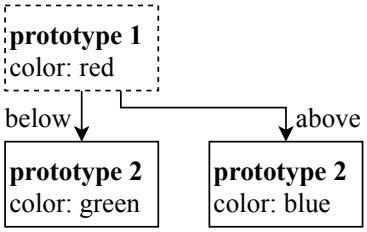


FIGURE 5.10: Query with multiple relations.

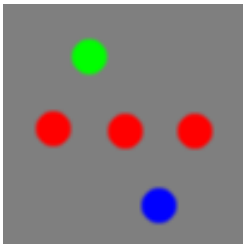


FIGURE 5.11: Scene for the query from Figure 5.10.

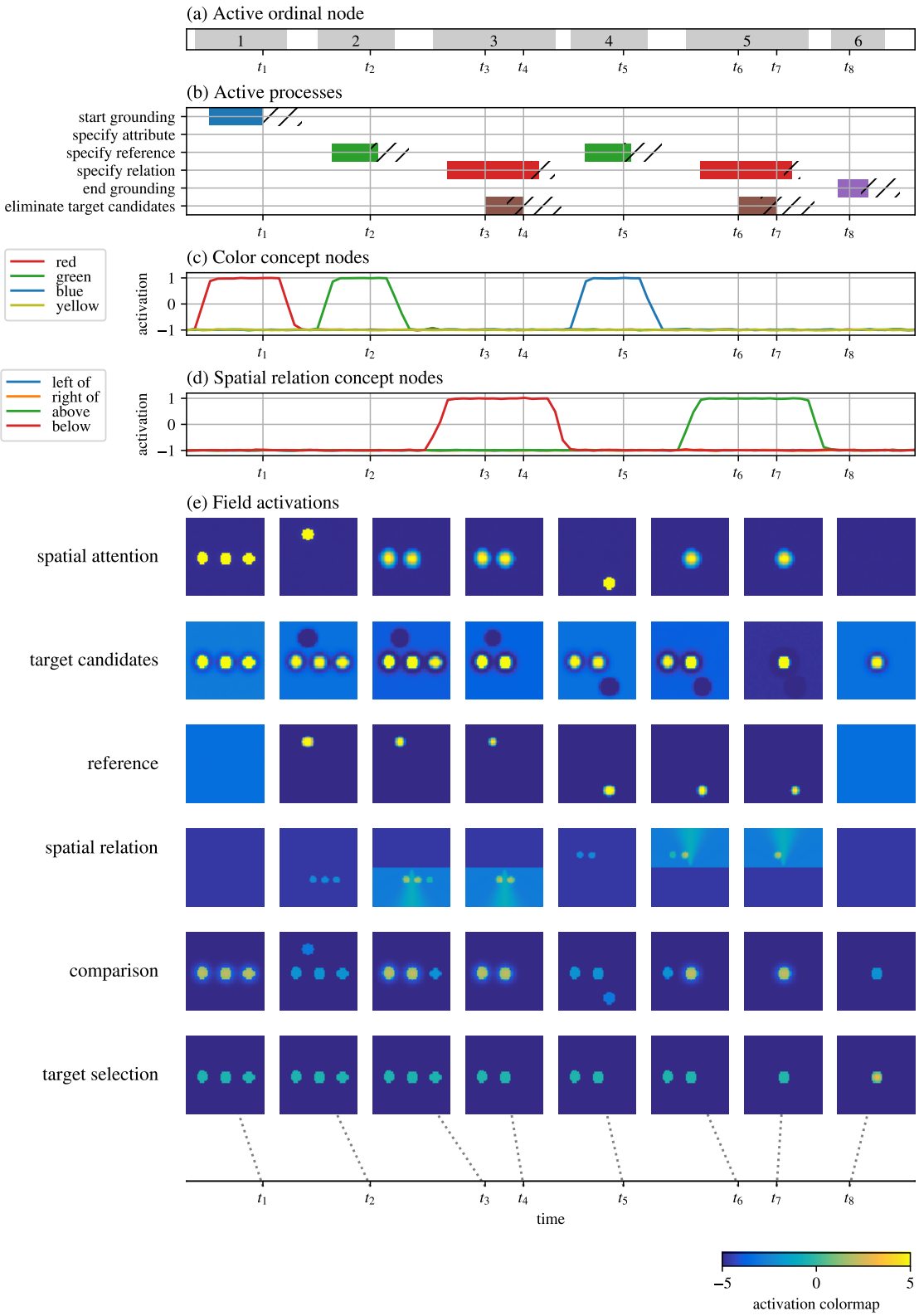


FIGURE 5.12: Time course of relevant parts of the architecture as it grounds the concept in Figure 5.10. **(a)** Active ordinal nodes. **(b)** Active processes. **(c)** Activation time courses of color concept nodes. **(d)** Activation time courses of spatial relation concept nodes. **(e)** Activation snapshots of relevant fields at time points of interest.

node, causing the spatial relation field to form peaks on the relative positions of the red target candidates that are below the green reference object.

At time t_4, the eliminate target candidates process intention has eliminated all target candidates that are not below the green object.

At time t_5, ordinal node 4 has activated the specify reference process intention and the blue color concept node, causing the blue object to be stored in the reference field.

At time t_6, ordinal node 5 has activated the specify relation process intention and the above spatial relation concept node, causing the spatial relation field to form peaks on the relative positions of the red target candidates that are above the blue reference object.

At time t_7, the eliminate target candidates process intention has eliminated all target candidates that are not above the blue object.

At time t_8, ordinal node 6 has activated the end grounding process intention, causing the remaining red target candidate to be selected in the target selection field.

5.5 Chaining relations

Consider the scene in Figure 5.14, and assume that a speaker wants to refer to the leftmost blue object. This object cannot be disambiguated based on its attributes alone, nor can it be disambiguated based on the fact that it is below a red object, since there are two red objects that have a blue object below it. However, it can be uniquely described by the denotational phrase "a blue object below a red object below a green object" (Figure 5.13).

In this scenario, we first have to find a red object below the green object, remember that red object, and then find a blue object that is below this previously selected red object. This is achieved by the following grounding strategy:

```
1 start grounding (color: red)
2 specify reference (color: green, source: scene)
3 specify relation (spatial relation: below)
4 end grounding
5 start grounding (color: blue)
6 specify reference (color: red, source: mental
    map)
7 specify relation (spatial relation: below)
```

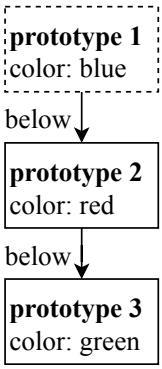


FIGURE 5.13: Query with chained relations.

FIGURE 5.14: Scene for the query from Figure 5.13.

```
8 end grounding
```

The red object assumes the role of the target object in the first grounding step, and the role of the reference object in the second grounding step. It is thus necessary to store the red object in the mental map after the first grounding step and refer to the stored red object in the second grounding step.

Figure 5.15 depicts the activation time course of relevant nodes and fields upon grounding the query from Figure 5.13. At time t_1, ordinal node 1 has activated the start grounding process intention and the red color concept node, which has caused the two red objects to be stored in the target candidates field.

At time t_2, ordinal node 2 has activated the specify reference process intention and the green color concept node, which has caused the green object to be stored in the reference field.

At time t_3, ordinal node 3 has activated the specify relation process intention and the below spatial relation concept node. By time t_4, this has caused all red target candidates which are not below the green object to be eliminated.

At time t_5, ordinal node 4 has activated the end grounding process intention, which has caused the target selection field to select the red object. As a result, this red object was stored in the mental map.

At time t_6, ordinal node 5 has activated the start grounding process intention and the blue color concept node, which has caused the target candidates field to form peaks on the three blue objects.

At time t_7, ordinal node 6 has activated the specify reference process intention, the red color concept node and the from mental map node. This has caused spatial attention to be directed to the red object from the mental map, which was stored in the reference field.

At time t_8, ordinal node 7 has activated the specify relation process intention and the below spatial relation concept node. By time t_9, this has caused all blue target candidates which are not below the red reference object to be eliminated.

At time t_{10}, ordinal node 8 has activated the end grounding process intention, which has caused the target selection field to select the blue target object.

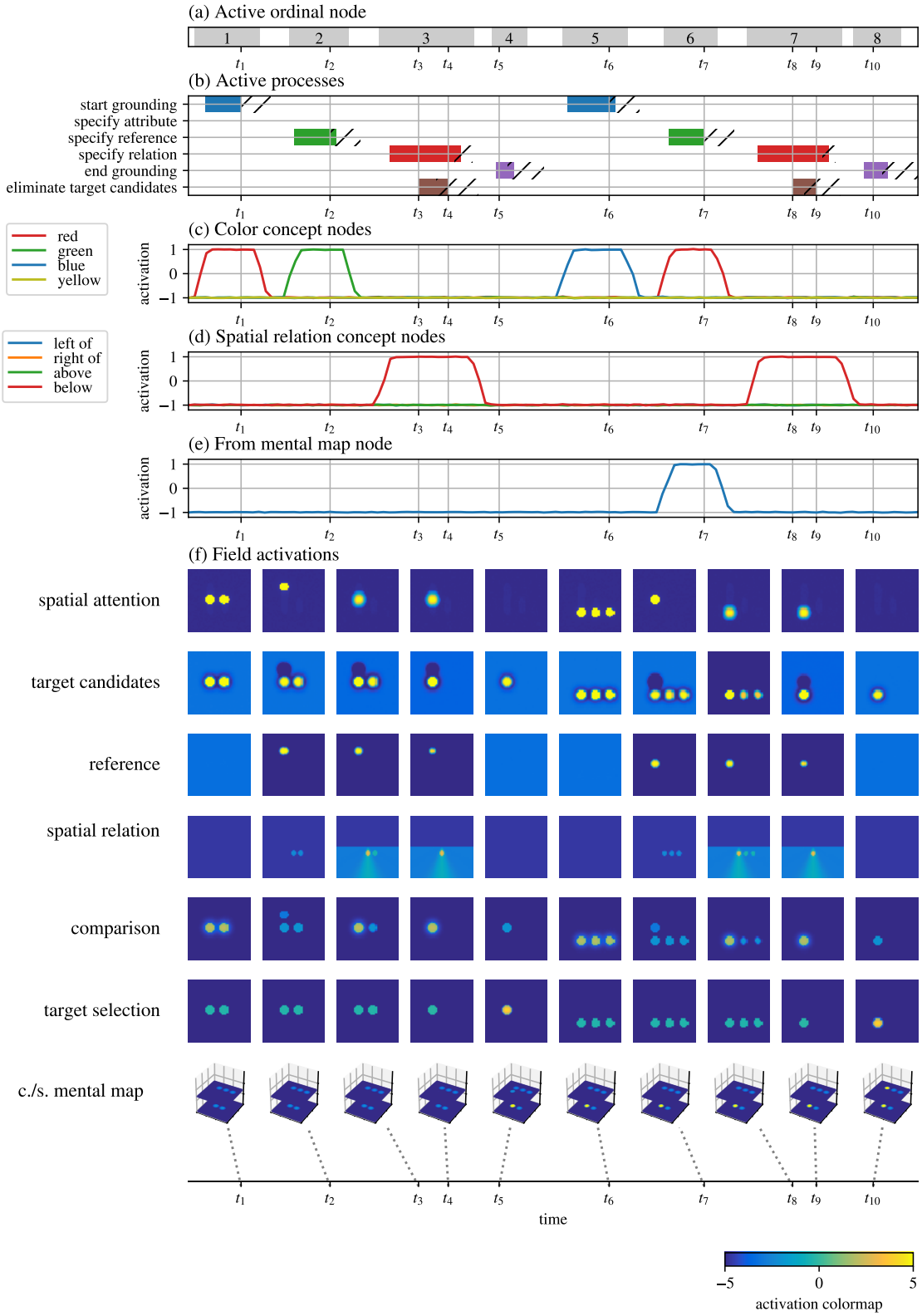


FIGURE 5.15: Time course of relevant parts of the architecture as it grounds the concept in Figure 5.13. **(a)** Active ordinal nodes. **(b)** Active processes. **(c)** Activation time courses of color concept nodes. **(d)** Activation time courses of spatial relation concept nodes. **(e)** Activation time course of from mental map node. **(f)** Activation snapshots of relevant fields at time points of interest.

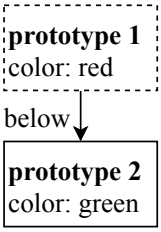


FIGURE 5.16: Query with multiple relations.

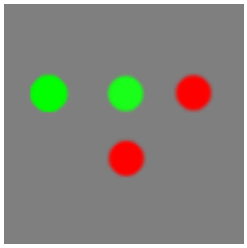


FIGURE 5.17: Scene for the query from Figure 5.16.

5.6 Single relation with backtracking

Consider again the query "a red object below a green object" (Figure 5.16), this time with the scene in Figure 5.17. Notice that there are two green objects. Thus, we cannot directly select a set of red target candidates and a green object from the scene, since the reference field might end up forming a peak on the wrong green object. Instead, we need to consider the selection of the green object as a grounding step of its own.

A grounding strategy for this scenario is given as follows:

```
1 start grounding (color: green)
2 end grounding
3 start grounding (color: red)
4 specify reference (color: green, source: mental
    map)
5 specify relation (spatial relation: below)
6 end grounding
```

First, the green object is selected as a target object, enforcing a selection decision. Then, a red object below the selected green object is selected. The backtracking machinery of the architecture, which consists of the no target candidates node and the inhibition of return field, ensures that in case a wrong green object is selected, the grounding starts again from the beginning and selects a different green object.

Figure 5.18 depicts the time course of relevant nodes and fields upon performing that grounding strategy. At time t_1, ordinal node 1 has activated the start grounding process intention and the green color concept node, causing the two green target candidates to be stored.

At time t_2, ordinal node 2 has activated the end grounding process intention, causing an arbitrary selection decision between the two green objects to be made in the target selection field. Note that the left green object is selected, which is the wrong choice. However, at this stage, the system is unable to tell that it was a wrong choice. As usual, the green object is stored in the mental map. Moreover, a peak has formed in the inhibition of return field on the position of the selected green object.

At time t_3, ordinal node 3 has activated the start grounding process intention and the red color concept node, which

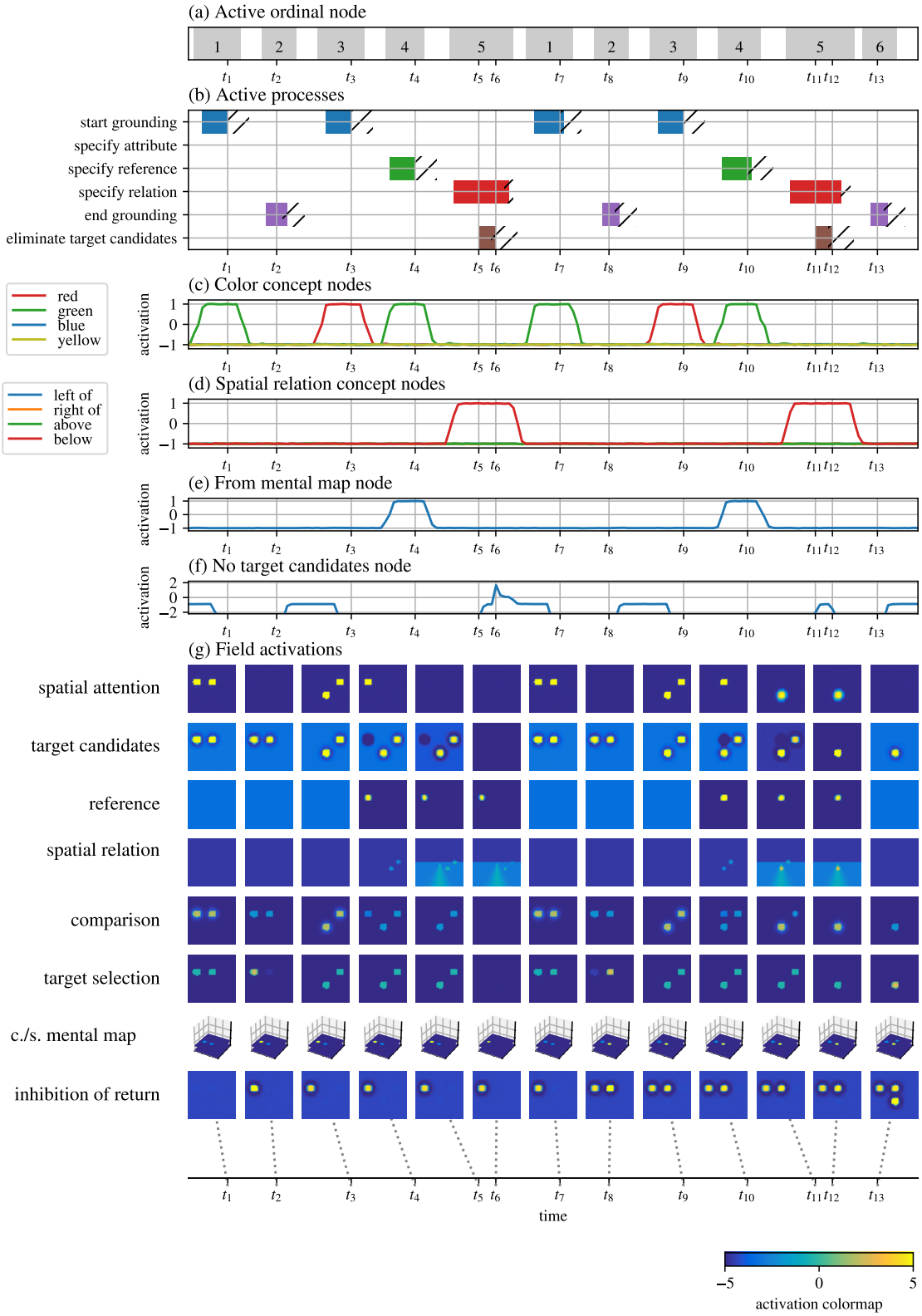


FIGURE 5.18: Time course of relevant parts of the architecture as it grounds the concept in Figure 5.16. **(a)** Active ordinal nodes. **(b)** Active processes. **(c)** Activation time courses of color concept nodes. **(d)** Activation time courses of spatial relation concept nodes. **(e)** Activation time course of from mental map node. **(f)** Activation time course of no target candidates node. **(g)** Activation snapshots of relevant fields at time points of interest.

has caused the two red objects to be stored as target candidates.

At time t_4, ordinal node 4 has activated the specify reference process intention, the green color concept node and the from mental map node, causing the reference field to form a peak on the previously selected green object from the mental map.

At time t_5, ordinal node 5 has activated the specify relation process intention and the below spatial relation concept node, causing the spatial below pattern to be fed into the spatial relation field. Since there are no red target candidates below the green reference object, no peaks form in that field and, consequently, spatial attention is not directed to any target candidates. The comparison field thus also does not form peaks on any target candidates.

At time t_6, inhibition from the eliminate target candidates process intention has eliminated all target candidates, since none of them received excitatory support from the comparison field. The fact that the target candidates field is now empty releases inhibition from the no target candidates node, causing it to become active. Subsequently, it inhibits all memory nodes, causing the sequence to start again from the beginning. Moreover, it inhibits the mental map, emptying it for the new grounding attempt.

At time t_7, ordinal node 1 has activated the start grounding process intention and the green color concept node, causing the two green target candidates to be stored.

At time t_8, ordinal node 2 has activated the end grounding process intention, causing a selection decision between the two green objects to be made in the target selection field. Due to inhibition from the inhibition of return field, the right green object is selected this time.

At time t_9, ordinal node 3 has activated the start grounding process intention and the red color concept node, which has caused the two red objects to be stored as target candidates.

At time t_{10}, ordinal node 4 has activated the specify reference process intention, the green color concept node and the from mental map node, causing the reference field to form a peak on the previously selected green object from the mental map.

At time t_{11}, ordinal node 5 has activated the specify relation process intention and the below spatial relation concept node, which by time t_{12} has caused all target candidates

which are not below the green reference object to be eliminated.

At time t_{13}, ordinal node 6 has activated the end grounding process intention, which has caused the remaining red target candidate to be selected in the target selection field.

Discussion 6

This master thesis introduced a neural process model that is able to ground combinatorial concepts in perception. Those parts of the architecture that have been present in previous architectures – the perceptual system, the attentional system, atomic concepts, apprehension of spatial relations, working memory representations, role-filler binding, and process organization – have been discussed in depth by Richter (2018). The discussion of the present architecture thus focuses on its novel contributions – the grounding strategy representation, the sequential elimination of target candidates, and the passing of parts of the activation state to future grounding processes via the mental map.

Richter, M. (2018). *A neural dynamic model for the perceptual grounding of spatial and movement relations.* (Doctoral dissertation, Bochum, Ruhr-Universität Bochum)

The discussion is organized as follows. Section 6.1 discusses arguments for the claim that combinatorial concepts are transformed into a sequential grounding strategy, which is then performed in sequence. Moreover, it motivates the choice of the instruction set. Section 6.2 discusses a range of models for how combinatorial structure may be represented in the human brain, and concludes that a sequential representation interfaces most naturally with the grounding system. Section 6.3 addresses the productivity, compositionality and systematicity challenge, which is widely believed to necessitate Turing-machine-like capacities in the brain, and shows that our architecture can account for these challenges while being significantly more restricted in its computational capacities than a Turing machine. Section 6.4 compares our grounding system to some other models for the perceptual grounding of con-

cepts. Section 6.5 addresses limitations of the architecture in accounting for all aspects of concept combination and suggests possible extensions to the architecture as future research directions.

6.1 The grounding strategy

Recall that the GSEnc transforms combinatorial concepts into a grounding strategy, i.e., a sequence of instructions that have to be performed in order to ground it. Two aspects of this view require motivation. First, we need to motivate why the frames of a combinatorial concept should be grounded in a sequence of grounding steps as opposed to in parallel. Second, we need to motivate why the intruction set is the way it is and, relatedly, why the grounding of each frame should proceed by a sequence of target candidate elimination steps.

6.1.1 Sequentiality arguments

There are theoretical and empirical reasons to suppose that the grounding of the parts of a combinatorial proceeds sequentially, one part at a time. For instance, the parts are usually interdependent due to their relations. Thus, grounding a frame may require that other frames on which it depends have already been grounded before, demanding a sequential grounding. If frames were grounded in parallel, candidates for each frame would have to be selected blindly, and subsequently the relationships with the objects selected for other frames would have to be verified. In case the relationships do not fit, a new blind selection of candidates for each frame would have to be made. This is an extremely inefficient way to ground a frame graph. It makes the unrealistic prediction that the time it takes to ground a frame graph consisting of k frames in a scene with n objects scales proportionally with the number of k-subsets of a set of n elements, i.e., $\frac{n!}{k!(n-k)!}$. In contrast, when a selection decision for one or more of the frames has already been made, other frames can be grounded significantly more efficiently by using the already grounded frames as reference. For instance, prototype 2 from Figure 2.3 may be grounded by finding all red objects that are both below the object selected for prototype 3 and above the object selected for

prototype 4, which allows to restrict the search space to a narrow range and to decrease the search time dramatically.

Moreover, it is likely that grounding each part requires certain neural resources that only exist once (e.g., spatial attentional resources and working memory of objects and relations). Using these neural resources to simultaneously ground two different parts would lead to interference, even if these parts are mutually independent.

This is backed by empirical evidence. Logan (1994) found that the time it takes to ground a relation between two objects (e.g., finding a red object above a blue object) increases proportionally with the number of distractors that are distinguished from the two target objects only by their spatial relation. This suggests that discriminating object pairs based on their spatial relation requires selective spatial attention, and that the consideration of different candidate object pairs proceeds sequentially. Franconeri, Scimeca, Roth, Helseth, and Kahn (2012) review further evidence that objects are attended to individually and sequentially in search tasks involving spatial relations.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1015

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227

Further support comes from the fact that language grounding usually proceeds in real time as a sequential linguistic representation is processed, i.e., people raise attention to the objects, one by one, as they are mentioned in an unfolding discourse (Tanenhaus et al., 1995).

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

6.1.2 Arguments for instruction set

The particular choice of instruction set presented in Chapter 3 requires motivation. Section 6.1.1 has already argued why the grounding of the frames plausibly proceeds sequentially. This motivates that the sequence of instructions is divided into blocks of separate frame grounding processes wrapped by a start grounding instruction and an end grounding instruction. However, this does not yet establish that the grounding of each individual frame should also proceed by a sequence of instructions that are executed one after another. It would be conceivable that the set of attributes and set of relations in a frame are processed in parallel.

Consider, first, the fact that the start grounding instruction takes a single attribute value as parameter and results in a selection of target candidate objects with that attribute value, and the fact that the target candidates are subsequently iteratively eliminated by a sequence of specify

attribute instructions. A conceivable alternative would be that the start grounding instruction takes all attribute values of the frame as parameters and results in a simultaneous consideration of all of them, selecting only objects as target candidates that have all the specified attribute values. Another conceivable alternative would be that the start grounding instruction takes no parameter, selecting all perceivable objects as target candidates, and subsequently a single specify attributes instruction with all attribute values as parameters eliminates all target candidates that do not have the specified attribute values. These alternatives are made implausible by psychophysical evidence. In a series of experiments on the concurrent discrimination of different attributes (color, form, and motion) reported by Lee, Koch, and Braun (1999), it was found that different discriminations draw on the same limited attentional capacities. Thus, attending to one attribute value comes at the expense of not being able to attend to another attribute value, even if they are values of different attributes. This makes it plausible that attention to the attribute values proceeds sequentially. Moreover, Burigo and Knoeferle (2015) review covert and overt attentional studies during spoken language comprehension. In these studies, it is found that upon processing a noun phrase, the words in that phrase are processed in an incremental fashion and constrain spatial attention to relevant target candidates: Initially, spatial attention is not directed. Upon hearing the first attribute value, spatial attention is divided between all objects with that attribute value. Subsequently, upon hearing each new attribute value, attention narrows down to all objects that have all attribute values mentioned so far. This gives support to the initial attribute-based attentional pop-out triggered by the start grounding instruction and the subsequent iterative elimination triggered by the specify attribute instructions.

Lee, D. K., Koch, C., & Braun, J. (1999). Attentional capacity is undifferentiated: Concurrent discrimination of form, color, and motion. *Perception & Psychophysics*, *61*(7), 1241–1255

Burigo, M. & Knoeferle, P. (2015). Visual attention during spatial language comprehension. *PLoS ONE*, *10*(1), e0115758. doi:10.1371/journal.pone.0115758

Consider, next, the fact that spatial relations are processed sequentially as opposed to in parallel. For example, the fact that prototype 2 from Figure 2.3 is both below prototype 3 and above prototype 4 is processed by first eliminating all target candidates that are not below prototype 3 and then eliminating all target candidates that are not above prototype 4 (Figure 3.8, lines 9-12). In principle, it is conceivable that all objects that are not below prototype 3 and all objects that are not above prototype 4

are eliminated in parallel. However, as found by Holcombe, Linares, and Vaziri-Pashkam (2011), applying spatial relations requires selective attention to the objects, which can only happen in sequence.

Holcombe, A. O., Linares, D., & Vaziri-Pashkam, M. (2011). Perceiving spatial relations via attentional tracking and shifting. *Current Biology*, *21*(13), 1135–1139

6.2 Representing combinatorial structures

Recall that a combinatorial concept, which is the output of a semantical analysis system as proposed by Jackendoff (2002), is transformed into a sequential grounding strategy by the GSEnc, which so far has been described at a functional rather than implementational level. We have not committed to how the combinatorial concept is represented prior to being transformed into a grounding strategy. Two alternatives come to mind. First, the combinatorial concept could be represented as a recursively nested symbolic structure. Second, it could be represented as an appropriately structured sequence, i.e., a sequence in which the recursive nesting is implicit (e.g., through bracketing). This section considers both possibilities in turn.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution.* Oxford University Press

6.2.1 Representing recursive structure explicitly

6.2.1.1 The LOTH view

The traditional model of representing recursive structure explicitly is that of a recursive data structure from object-oriented programming, as proposed by the LOTH: Objects have attributes, which can be objects again that in turn have their own attributes, etc. At the implementational level, this is achieved by pointers to arbitrary positions in memory, which in turn contain pointers to other positions in memory, etc. Problematically, neural populations cannot represent pointers to arbitrary other neural populations in this way, since synaptic connections among populations are fixed on short time scales. Thus, a population that is able to represent pointers to arbitrary other neural populations would have to be connected to each of these populations, which is unrealistic. Hence, the LOTH model of representing combinatorial structure gives us no insights regarding neural realization.

6.2.1.2 The LISA model

Hummel, J. E. & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220–264

Hummel and Holyoak (2003) propose the *LISA* model for how recursively nested symbolic structures can be neurally represented. Each part of a combinatorial concept is represented by a population of neurons that are organized along four levels of a hierarchy.

The first level consists of a set of neural populations representing *subsymbols* like MALE, FEMALE, ADULT, HUMAN, etc.

The second level consists of a set of neural populations representing *symbols*, which are connected to a set subsymbols that define their semantics. For instance, the symbol BILL is connected to the subsymbol populations for MALE, ADULT, HUMAN, etc. Similarly, the symbol LOVER is connected to the subsymbol populations for HAS-EMOTION, EMOTION-POSITIVE, etc.

The third level consists of a set of neural populations representing *subpropositions* whose semantics are defined by virtue of their connections to the symbols. For instance, the role-binding BILL+LOVER is represented as a neural population that is connected with the symbols BILL and LOVER.

The fourth level consists of a set of neural populations representing *propositions*, which are connected to a set of subpropositions that define their semantics. For instance, the proposition LOVES(BILL, MARY) is represented by its connection with the subpropositions BILL+LOVER and MARY+BELOVED.

A downside of the LISA model is that it is limited to structures of a certain depth. While there may be a realistic cognitive limit on the depth of the structures that humans are able to represent, it is unlikely that this depth limit occurs at the level of single relations.

Stewart, T. & Eliasmith, C. (2012). Compositionality and biologically plausible models. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press

A more serious problem with the LISA model is the fact that each recursively nested structure is represented by its own neural population. Stewart and Eliasmith (2012) estimate that being able to represent every proposition of the form RELATION(AGENT, THEME) with LISA requires an unrealistic 30 billion neural populations, while the human brain only contains around 100 billion neurons. If higher-order propositions like KNOW(RELATION(AGENT, THEME)) were desired as well, the number would explode even further.

Alternatively, if new propositions were to be built on

the fly, this would require altering synaptic connections, which is likely to take longer than the time required to read and understand a novel linguistic expression.

6.2.1.3 The Neural Blackboard Architecture

Van der Velde and De Kamps (2006) propose the *Neural Blackboard Architecture*. This architecture is somewhat similar to the LISA model, but instead of representing each filler-role binding by its own neural population, it instead allows to flexibly assign roles to fillers via gating neurons.

Van der Velde, F. & De Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, *29*(1), 37–70

There is a neural population for each noun, and a fixed number of neural populations called *noun assemblies*. Each noun assembly is connected to every noun, but these connections are gated. Thus, each noun assembly can be flexibly bound to an arbitrary noun by "opening the gate" of that connection. An analogous mechanisms exists for the other word types in the form of *verb assemblies*, *adjective assemblies*, *preposition assemblies*, *clause assemblies*, etc.

Each noun assembly is further connected to a set of *role assemblies* (e.g., *agent assemblies*, *theme assemblies*, etc). Again, these connections are gated, allowing each noun assembly to be bound to an arbitrary role assembly. The role assemblies can in turn be combined to relational assemblies. Again, analogous mechanisms exist for the other word types.

Stewart and Eliasmith (2012) describe some problems with this architecture. First, they estimate that this architecture requires around 800 million neurons, which is significantly less than the number of neurons required by the LISA model, but still rather much. Second, they note that the complete connectivity between noun assemblies and nouns requires long-distance complete connectivity, which is at odds with the local and sparse connectivity that is actually observed in the cortex. Third, they note that the architecture does not feature graceful degradation. Rather, damage to small parts of the architecture can lead to patterns of failure that are not observed in human language behavior.

Stewart, T. & Eliasmith, C. (2012). Compositionality and biologically plausible models. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press

6.2.1.4 Vector-Symbolic Architectures

Vector-Symbolic Architectures (VSAs) (e.g., Smolensky, 1990; Plate, 1995) represent atomic concepts as high-dimensional

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1-2), 159–217

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*(3), 623–641

vectors and define a range of algebraic operations that allow to combine vectors representing atomic concepts into tensors or vectors representing combinatorial concepts. These combinatorial concepts can in turn be combined into yet more complex combinatorial concepts. Inverse algebraic operations allow to reconstruct the original component concepts from the combinatorial concepts.

Stewart, T. & Eliasmith, C. (2012). Compositionality and biologically plausible models. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford Handbook of Compositionality.* Oxford: Oxford University Press

While early work on VSAs took these vectors or tensors to be activation values in connectionist networks, it has by now been recognized that this is biologically implausible. However, Stewart and Eliasmith (2012) demonstrate how VSAs can be neurally realized with the *Neural Engineering Framework* (NEF).

The NEF may be criticized for employing optimization methods which are so powerful that they can learn virtually any mathematical operation if it can in theory be implemented through networks of neurons. While the resulting model is then built out of biologically realistic neurons, it is questionable whether the model as a whole is also biologically realistic. Moreover, the algebraic operations that allow to combine atomic concept vectors into combinatorial concept tensors or vectors are to a large part contrived and no reason is given for why the brain should employ them.

6.2.2 Representing recursive structure implicitly in a sequence

An alternative approach to representing the recursive structure of a combinatorial concept explicitly is to represent it by means of an appropriately structured symbol sequence. "Appropriately structured" means that the sequence is built according to rules of syntax in such a way that information about how the symbols are recursively nested is implicit in the sequence and can in principle be extracted from it.

As an example, consider again the combinatorial concept from Figure 2.3, i.e., "a red object right of a red object below a green diagonal object and above a blue object". A sequential representation of that concept might look as follows:

```
1 (red rightOf (red below (green diagonal) above
    (blue)))
```

The way the brackets are placed implicitly determines the recursive structure. This becomes clear when adding indentation as follows:

```
(
    red
    rightOf (
        red
        below (
           green
           diagonal
        )
        above (
            blue
        )
    )
)
```

The neural systems that work with such a representation merely need to process it in accordance with the rules of syntax, which ensures that the recursive structure is respected correctly.

Representing combinatorial structures as a sequence requires significantly fewer neural resources than either of the architectures considered so far. As described in Section 4.4, if synaptic connections are fixed, then it merely requires to repeat the set of concept nodes once for each ordinal node in a sequence and adding a gating mechanism for each of these copies. If synaptic connections are entrained by fast plasticity, then it only requires a single node for each concept.

Moreover, representing the combinatorial structure explicitly adds a lot of complexity but does not appear to serve any purpose. The only purpose of representing the combinatorial structure explicitly is to be able to access the parts of the combinatorial structure independently and in parallel. If we assume that conceptual processing proceeds sequentially, then there is no need for this, and thus it would be a waste of neural resources. In particular, upon processing a sentence, that sequential sentence would have to be transformed into a combinatorial concept whose combinatorial structure is explicitly represented, only to be transformed back into a sequence again upon grounding it.

In summary, our architecture can as of now remain agnostic as to how conceptual structure is represented prior

to being transformed into a grounding strategy, albeit a sequential representation interfaces most naturally with the sequential grounding strategy representation.

6.3 Addressing productivity, systematicity and compositionality

Recall from Section 2.3 that an important argument in favor of the CCTM has been its ability to account for the productivity, compositionality and systematicity of thought and language. The commonality in these arguments is that they show that the mind must operate on structured representations, that its operations must be sensitive to the combinatorial structure of those representations, and that there must be a homogeneous mechanism that processes combinatorial structures, regardless of what the parts of those structures are and how they are combined. These are properties exhibited by Turing machines, which Fodor and Pylyshyn take to show that the mind must be a Turing machine.

Our architecture can meet all of these challenges. It exhibits productivity, since it is able to ground an indefinite range of concepts by finite means. This is achieved by operating on grounding strategies which are structured representations, i.e., they are generated recursively out of parts.

The architecture also exhibits compositionality, since the meaning (i.e., denotation) of each concept is determined by the meanings (i.e., denotations) of its parts (i.e., frames, attributes and relations) and the way they are put together (i.e., the way the sequence of grounding processes is structured). This is achieved because the GSEx is sensitive to the combinatorial structure of the grounding strategy. The components of a combinatorial concept are sequentially grounded, while the state of the grounding system is passed on from one grounding step to the next grounding step through self-sustained fields. In particular, the mental map, which stores the objects that have been grounded thus far, allows future grounding processes to refer back to those objects. Through a chain of back-references of this kind, a hierarchical dependency structure between grounding

processes emerges, which ensures that the combinatorial structure of the concept is respected correctly. This sequential grounding and passing along of the state is to be distinguished from CCTM architectures that achieve productivity and compositionality by a recursive pattern of function calls.

Finally, the architecture exhibits systematicity, since the ability to ground some concepts is systematically related to the ability to ground other concepts. For example, the ability to ground "a green object below a red object" is systematically related to the ability to ground "a red object right of a red object below a green diagonal object and above a blue object". The grounding of these concepts employs the same cognitive mechanism. The only difference is the way the grounding strategy guides the grounding strategy executor, i.e., the way the grounding strategy is structured and the way that slots are filled with values.

To summarize, our architecture operates on structured representations, its operations are sensitive to the combinatorial structure of those representations, and there is a homogeneous mechanism that processes these combinatorial structures, regardless of what the parts of those structures are and how they are combined. This accounts for Fodor and Pylyshyn's arguments and shows that they can be met by a neurally plausible architecture that is significantly more restrained than a Turing machine.

6.4 Contrast to other approaches for the grounding of combinatorial concepts

SHRDLU (Winograd, 1971), the first computational model that was able to find referents for combinatorial concepts in the world, had access to an amodal symbolic representation of the world and matched a given combinatorial concept to that representation by a logical constraint satisfaction procedure. The system lacked a mechanism to build a scene representation out of perceptual input. Instead, it had prior access to a symbolic representation of the scene.

SAM (Brown et al., 1992) and *Ubiquitous Talker* (Nagao & Rekimoto, 1995) are more comprehensive models. Their amodal representation of the scene (the "knowledge

Winograd, T. (1971). *Procedures as a representation for data in a computer program for understanding natural language.* Massachusetts Institute of Technology, Project MAC

Brown, M. K., Buntschuh, B. M., & Wilpon, J. G. (1992). Sam: A perceptive spoken language-understanding robot. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*(6), 1390–1402

Nagao, K. & Rekimoto, J. (1995). Ubiquitous talker: Spoken language interaction with real world objects. *arXiv preprint cmp-lg/9505038*

base"), which they use to find referents for combinatorial concepts, is built from perceptual input. Moreover, they employ a natural language parsing framework, which yields a combinatorial syntactical representation of a linguistic expression, and then evaluate the syntactical representation through recursive function calls in order to find a referent in their knowledge base.

While being landmark works in AI which showed the field's potential and were impressive for their time, these types of architectures are unlikely to lead to insights regarding language and concept grounding in biological cognitive systems. Apart from not being neural models, they all find objects that match a given combinatorial concept by a form of amodal symbolic constraint satisfaction procedure. For instance, when finding "a red object above a green object", they do so by finding an amodal proposition in their knowledge base which states that two objects in the scene bear that spatial relation to each other. This is at odds with the finding that apprehending spatial relations is a perceptual mechanism that requires spatial attention (Logan, 1994). Moreover, matching combinatorial concepts to a symbolic representation of the scene rather than the scene itself loses a lot of the subtle visual information that could be put to use when finding a referent for a concept.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1015

Gorniak and Roy (2004) provide a more realistic contribution that is closer to our aims – an algorithmic model that grounds combinatorial concepts in perception. Like SAM and Ubiquitous Talker, their model parses a linguistic phrase into a combinatorial syntactical representation. However, it then evaluates this representation against a visual/geometric representation of the scene as opposed to an amodal knowledge base, finding a matching object. This model is able to cover an impressive range of tasks, accounts for empirical psychological data, and is committed to the GC stance. However, it incorporates representations like recursively nested data structures, which, as argued before, are difficult to realize neurally. Moreover, it operates on these nested data structures through recursive function calls that necessitate a call stack, which is also neurally implausible. This is in stark contrast to the more parsimonious approach of processing the parts of a combinatorial concept sequentially and passing on a minimal representation of the output of each grounding step to subsequent grounding steps.

Gorniak, P. & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470

Richter (2018, p. 12) reviews a range of other models for the grounding of concepts, all of which are lacking in one or more of the following respects:

Richter, M. (2018). *A neural dynamic model for the perceptual grounding of spatial and movement relations.* (Doctoral dissertation, Bochum, Ruhr-Universität Bochum)

1. They are algorithmic rather than neural, or at least include algorithmic building blocks.

2. They cannot deal with arbitrarily nested combinatorial concepts, but only with either atomic concepts or single relations.

3. They find matching objects against an amodal symbolic knowledge base, rather than grounding them in perception.

Thus, to the best of my knowledge, our architecture is the first neural process model for the perceptual grounding of arbitrarily nested combinatorial concepts.

6.5 Limitations and future research directions

Research on concept representation and composition still has a long way to go, making a comprehensive neural theory of these feats a distant goal (Barsalou, 2017). Nevertheless, progress in that direction can be made based on what is already known.

As of now, our architecture is limited to the perceptual spaces color, orientation, shape, and to spatial relations. To bring it closer toward a neural theory of concepts in general, an important step is therefore to extend it by additional attributes and relations. Additional attributes could be other perceptual attributes like texture or size, or more abstract conceptual attributes like age, speed, ripeness, material, force, hardness, arousal, etc. For the latter, an interesting research direction is to investigate how these abstract attributes may ultimately be grounded in perceptual attributes.

Another step that would bring the architecture closer toward a comprehensive neural theory of concept grounding is the ability to define high-level concept nodes in terms of low-level frame graphs – e.g., a tree is a green object above a vertical brown rectangle. Thus, activating high-level concept nodes causes these high-level concepts to be

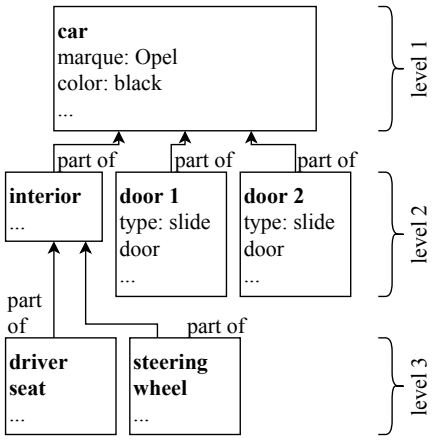


FIGURE 6.1: A frame graph for the combinatorial concept of a car composed of a hierarchy of parts.

grounded by virtue of initiating an appropriate lower-level grounding sequence. This mechanism should be hierarchical in nature, such that high-level concepts can be defined in terms of other high-level concepts – e.g., a forest is a tree to the left of a tree to the left of a tree. Moreover, it should allow for concept inheritance – e.g., a fir tree is a tree whose treetop has triangular shape.

Furthermore, it is clear that most human concepts, both atomic and combinatorial, do not have sharp boundaries, but are vague (e.g., Keefe and Smith, 1996). Rather than judging whether or not a given concept is applicable, humans instead seem to determine a degree of membership and adapt their behavior based on how certain they are that a given concept is applicable. In contrast, our model only tries to ground a given combinatorial concept without determining a degree of membership. Evaluating goodness of fit of combinatorial concepts is therefore a possible future extension to the architecture.

In addition, different feature dimensions seem to have different importance in establishing the degree of concept membership, which is usually modeled by assigning a weight to each feature dimension (e.g., Rosch and Mervis, 1975). Our architecture, in contrast, treats all feature dimensions to be of equal importance. In addition, the involvement of the various feature dimensions is different when activating a concept bottom-up as opposed to top-down. For example, encountering a penguin is unlikely to activate the concept of bird, since it does not exhibit typical bird features, but a top-down grounding process might still establish a penguin as a perfectly good member of the bird concept (Hampton, 2007). Adding the possibility to add different degrees of importance to feature dimensions is therefore a conceivable extension.

Another as of yet disregarded fact is that combinatorial concepts are not always grounded completely upon each encounter with an object falling under the concept. Instead, parts of the combinatorial concept may get grounded according to task demands. Especially for hierarchically nested combinatorial concepts, e.g., ones that represent an object as composed of a hierarchy of parts (Figure 6.1), the deep structure is not always considered completely. Instead, superficial information is activated more reliably across situations, whereas deeper information may be activated later when task demands require it (Forbus, Gentner, &

Law, 1995; Blanchette & Dunbar, 2000). For instance, upon encountering a car, it may first be identified by one or more characteristic features. When trying to enter the car, deeper levels of the car concept may get activated, e.g., representations of the doors and interior as parts of the car. After entering a car, still deeper levels of the hierarchy may get activated, e.g., representations of the driver seat and steering wheel. In contrast, our architecture always tries to ground a combinatorial concept completely, taking each of its parts into account. A possible future extension is to allow parts of the combinatorial concepts to get activated based on task demands.

Forbus, K. D., Gentner, D., & Law, K. (1995). Mac/fac: A model of similarity-based retrieval. *Cognitive Science*, *19*(2), 141–205

Blanchette, I. & Dunbar, K. (2000). How analogies are generated: The roles of structural and superficial similarity. *Memory & Cognition*, *28*(1), 108–124

It has furthermore been argued that the meanings of concepts change depending on how they are put together into combinatorial concepts. In contrast, our model can only handle intersections or unions of concepts, without altering the meaning of a concept based on the combinatorial concept in which it occurs. Gärdenfors (2014) gives the examples of adjective-noun combinations like "white skin", which is actually pinkish, "black skin", which is actually brown, and "large squirrel", which is not a large animal. These examples can be accommodated by regarding the adjectives to attain their meaning only in relation to the noun with which they are combined. The question of how the meanings of atomic concepts change depending on linguistic context has been investigated by semanticists at a functional level. A possible research direction is to find neural models for these context-dependent meaning changes.

Not only immediate linguistic context, but also discourse context (Werning & Cosentino, 2017) and visual context (Gorniak & Roy, 2004) can change the meanings of atomic and combinatorial concepts. For instance, Gorniak and Roy (2004) find that the word "middle" can have four different meanings depending on different visual contexts to be understood. Similarly, expressions like "the leftmost green object" may either denote an object which is the leftmost object and green, or an object which is the leftmost among the green objects. This is a challenging aspect of language and likely to require a great deal of research effort.

Werning, M. & Cosentino, E. (2017). The interaction of bayesian pragmatics and lexical semantics in linguistic interpretation: Using event-related potentials to investigate hearers' probabilistic predictions. In G. Gunzelman et al. (Eds.), *Proceedings of the 39th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society

Gorniak, P. & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470

The presented architecture so far only accounts for one direction of the grounding process – namely, grounding a previously activated combinatorial concept in perception. Future extensions could implement the converse process of describing a scene. This may involve devising a way to

activate existing high-level concepts from a large collection of concepts heuristically (e.g., as proposed by many, through typical features that are neither necessary nor sufficient for concept membership), and then making a final decision by trying to ground the activated concept. It may also involve enabling the system to combine concepts by itself.

Finally, our architecture lacks syntactical parsing of natural language, its transformation into a conceptual structure, and the subsequent transformation into a grounding strategy. While our architecture can remain agnostic as to how these processes are neurally realized and still comprise a relevant contribution to the perceptual grounding of combinatorial concepts, these aspects have to be addressed at some point if the goal is to create a comprehensive model of natural language grounding.

Conclusion 7

This master thesis introduced a neural process model based on Dynamic Field Theory that is able to ground combinatorial concepts describable by frame graphs in perception. Combinatorial concepts are transformed into a sequential grounding strategy by the GSEnc. The grounding strategy is then fed into the GSEx, which performs that grounding strategy, effectively raising attention to an object in the perceptual array that matches the combinatorial concept. In doing so, it passes parts of its activation state representing the outcome of one grounding step to subsequent grounding steps by virtue of the self-sustained mental map. This mental map allows grounding steps to refer back to the outcomes of previous grounding steps. Through a chain of back-references of this kind, a hierarchical dependency structure between grounding processes emerges, which mirrors the dependency structure between the parts of the combinatorial concepts.

The main contribution of this thesis is an extension of previous neural architectures for the grounding of attributes and relations (Lipinski et al., 2012; Richter et al., 2014; Richter et al., 2017; Richter, 2018) by the ability to ground arbitrarily nested combinatorial concepts.

The capabilities of the model were demonstrated in a set of 6 simulations with qualitatively different combinatorial concepts and scenes. The model succeeded in grounding the given combinatorial concepts in all simulations, demonstrating that all of these qualitatively different kinds of combinatorial concepts can be grounded by a single ar-

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neuro-behavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *38*(6), 1490–1511

Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., & Schöner, G. (2014). Autonomous neural dynamics to test hypotheses in a model of spatial language. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 2847–2852). Austin, TX: Cognitive Science Society

Richter, M., Lins, J., & Schöner, G. (2017). A neural dynamic model generates descriptions of object-oriented actions. *Topics in Cognitive Science*, *9*, 35–47

Richter, M. (2018). *A neural dynamic model for the perceptual grounding of spatial and movement relations.* (Doctoral dissertation, Bochum, Ruhr-Universität Bochum)

chitecture with a single set of parameters. Since other combinatorial concepts describable by frame graphs are not qualitatively different to the probed scenarios, there is reason to believe that the model is able to successfully ground arbitrary frame graphs.

A second contribution of this thesis is an embedding of the neural theory of atomic and combinatorial concepts underlying our grounding system in the literature of psychological theories of concepts. It was argued that the atomic concepts introduced by previous architectures are committed to a prototype theory of concepts, which represents concepts as a probability distribution in a conceptual or perceptual space. Moreover, it was demonstrated that the class of combinatorial concepts that our architecture is able to ground are all those concepts that can be described as frame graphs.

A third contribution of this thesis is a clear interface between the grounding system and language processing systems. It was argued that the grounding system can be conceived of as an extension to the Parallel Architecture (Jackendoff, 2002) that takes the output of the semantical/conceptual analysis system as input. As such, it can be regarded as an account for the grounding of complex linguistic expressions, while being agnostic to how the phonological, syntactical and semantical analysis systems are neurally realized.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution.* Oxford University Press

All in all, this thesis contributes another building block that brings DFT closer towards comprehensive models of the higher cognitive feats. It offers interfaces to integrate it with other work by the DFT research community on object recognition, scene representation, grounding movement relations, mental models, motor control, and more. Together, these architectures already comprise a striking model of important parts of the human brain. When extended in the ways suggested in the previous chapter, these architectures may be lifted towards realistic models of reasoning and language and, ultimately, all of cognition.

Bibliography

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211–227.

Barsalou, L. W. (1992). Frames, concepts, and conceptual fields. In A. Lehrer & E. F. Kittay (Eds.), *Frames, fields, and contrasts: New essays in lexical and semantic organization*. Lawrence Erlbaum Associates, Inc.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Barsalou, L. W. (2016). On staying grounded and avoiding quixotic dead ends. *Psychonomic Bulletin & Review*, *23*(4), 1122–1142.

Barsalou, L. W. (2017). Cognitively plausible theories of concept composition. In J. A. Hampton & Y. Winter (Eds.), *Compositionality and concepts in linguistics and psychology* (pp. 9–30). Springer International Publishing.

Blanchette, I. & Dunbar, K. (2000). How analogies are generated: The roles of structural and superficial similarity. *Memory & Cognition*, *28*(1), 108–124.

Brown, M. K., Buntschuh, B. M., & Wilpon, J. G. (1992). Sam: A perceptive spoken language-understanding robot. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*(6), 1390–1402.

Burigo, M. & Knoeferle, P. (2011). Visual attention during spatial language comprehension: Is a referential linking hypothesis enough? In L. Carlson, C. H. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Burigo, M. & Knoeferle, P. (2015). Visual attention during spatial language comprehension. *PLoS ONE*, *10*(1), e0115758. doi:10.1371/journal.pone.0115758

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.

Cohen, B. & Murphy, G. L. (1984). Models of concepts. *Cognitive Science*, *8*(1), 27–58.

Fodor, J. A. (1975). *The language of thought*. New York, NY: Crowell.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Forbus, K. D., Gentner, D., & Law, K. (1995). Mac/fac: A model of similarity-based retrieval. *Cognitive Science*, *19*(2), 141–205.

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge, MA: MIT Press.

Georgopoulos, A. P., Kettner, R. E., & Schwartz, A. B. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. *Journal of Neuroscience*, *8*(8), 2928–2937.

Glaser, W. R. (1992). Picture naming. *Cognition*, *42*(1-3), 61–105.

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.

Gorniak, P. & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science*, *31*(3), 355–384.

Hampton, J. A. & Winter, Y. (2017). *Compositionality and concepts in linguistics and psychology*. Springer International Publishing.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*, 335–346.

Henson, R. & Burgess, N. (1997). Representations of serial order. In *4th neural computation and psychology workshop, london, 9–11 april 1997* (pp. 283–300). Springer.

Holcombe, A. O., Linares, D., & Vaziri-Pashkam, M. (2011). Perceiving spatial relations via attentional tracking and shifting. *Current Biology*, *21*(13), 1135–1139.

Hubel, D. H. & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591.

Hummel, J. E. & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220–264.

Jackendoff, R. (1987). The status of thematic relations in linguistic theory. *Linguistic Inquiry*, *18*(3), 369–411.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

Jancke, D., Erlhagen, W., Dinse, H. R., Akhavan, A. C., Giese, M., Steinhage, A., & Schöner, G. (1999). Parametric population representation of retinal location: Neuronal interaction dynamics in cat primary visual cortex. *Journal of Neuroscience*, *19*(20), 9016–9028.

Janssen, T. M. et al. (2012). Compositionality: Its historic context. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The oxford handbook of compositionality* (pp. 19–46).

Keefe, R. & Smith, P. (1996). *Vagueness: A reader*. Cambridge, MA: MIT Press.

Keil, F. C. & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 221–236.

Kounatidou, P., Richter, M., & Schöner, G. (2018). A neural dynamic architecture that autonomously builds mental models. In C. Kalish, M. A. Rau, X. Zhu, & T. T. Rogers (Eds.), *Proceedings of the 40th annual meeting of the cognitive science society*. Austin, TX: Cognitive Science Society.

Lee, C., Rohrer, W. H., & Sparks, D. L. (1988). Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature*, *332*(6162), 357.

Lee, D. K., Koch, C., & Braun, J. (1999). Attentional capacity is undifferentiated: Concurrent discrimination of

form, color, and motion. *Perception & Psychophysics*, *61*(7), 1241–1255.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neuro-behavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *38*(6), 1490–1511.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1015.

Logan, G. D. & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (Chap. 13). Cambridge, MA: MIT Press.

Lomp, O., Zibner, S. K. U., Richter, M., Ranó, I., & Schöner, G. (2013). A software framework for cognition, embodiment, dynamics, and autonomy in robotics: Cedar. In *International conference on artificial neural networks* (pp. 475–482). Springer.

Ludlow, P. (2018). Descriptions. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Fall 2018). Metaphysics Research Lab, Stanford University.

Machery, E. (2009). *Doing without concepts*. Oxford University Press.

Margolis, E. & Laurence, S. (2019). Concepts. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2019). Metaphysics Research Lab, Stanford University.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: WH Freeman.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Minsky, M. (1977). A framework for representing knowledge. In P. Winston (Ed.), *The psychology of computer vision*. New York, NY: McGraw-Hill.

Murata, A., Gallese, V., Luppino, G., Kaseda, M., & Sakata, H. (2000). Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP. *Journal of Neurophysiology*, *83*(5), 2580–2601.

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Nagao, K. & Rekimoto, J. (1995). Ubiquitous talker: Spoken language interaction with real world objects. *arXiv preprint cmp-lg/9505038*.

Nandy, A. S., Sharpee, T. O., Reynolds, J. H., & Mitchell, J. F. (2013). The fine structure of shape tuning in area V4. *Neuron*, *78*(6), 1102–1115.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Petersen, W. (2007). Representation of concepts as frames. *The Baltic International Yearbook of Cognition, Logic and Communication*, *2*, 151–170.

Pitt, D. (2018). Mental representation. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2018). Metaphysics Research Lab, Stanford University.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*(3), 623–641.

Port, R. F. & Van Gelder, T. (1995). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT press.

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, *22*(2), 253–336.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(July), 576–582.

Regier, T. & Carlson, L. A. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, *130*(2), 273.

Rescorla, M. (2017). The computational theory of mind. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2017). Metaphysics Research Lab, Stanford University.

Richter, M. (2018). *A neural dynamic model for the perceptual grounding of spatial and movement relations.* (Doctoral dissertation, Bochum, Ruhr-Universität Bochum).

Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., & Schöner, G. (2014). Autonomous neural dynamics to test hypotheses in a model of spatial language. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual meeting of the*

cognitive science society (pp. 2847–2852). Austin, TX: Cognitive Science Society.

Richter, M., Lins, J., & Schöner, G. (2017). A neural dynamic model generates descriptions of object-oriented actions. *Topics in Cognitive Science*, *9*, 35–47.

Richter, M., Sandamirskaya, Y., & Schöner, G. (2012). A robotic architecture for action selection and behavioral organization inspired by human cognition. In *2012 IEEE/RSJ international conference on intelligent robots and systems* (pp. 2457–2464). New York, NY: IEEE.

Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Sandamirskaya, Y. & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179.

Schneegans, S. & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics*, *106*(2), 89–109.

Schneegans, S., Spencer, J., & Schöner, G. (2015). Integrating "what" and "where": Visual working memory for objects in a scene. In G. Schöner & J. Spencer (Eds.), *Dynamic thinking: A primer on dynamic field theory*. New York, NY: Oxford University Press.

Schöner, G. & Spencer, J. (2015). *Dynamic thinking: A primer on dynamic field theory*. New York, NY: Oxford University Press.

Shapiro, L. (2010). *Embodied cognition*. Routledge.

Smith, E. E. (1988). Concepts and thought. *The Psychology of Human Thought*, *147*.

Smith, E. E. & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1-2), 159–217.

Sowa, J. F. et al. (2000). *Knowledge representation: Logical, philosophical, and computational foundations*. Pacific Grove, CA: Brooks/Cole.

Stewart, T. & Eliasmith, C. (2012). Compositionality and biologically plausible models. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford Handbook of Compositionality*. Oxford: Oxford University 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.

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*(2), 230–265.

Van der Velde, F. & De Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, *29*(1), 37–70.

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and brain sciences*, *21*(5), 615–628.

Werning, M. (2005). Right and wrong reasons for compositionality. *The Compositionality of Meaning and Content: Foundational Issues*, *1*, 285–309.

Werning, M. & Cosentino, E. (2017). The interaction of bayesian pragmatics and lexical semantics in linguistic interpretation: Using event-related potentials to investigate hearers' probabilistic predictions. In G. Gunzelman et al. (Eds.), *Proceedings of the 39th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Winograd, T. (1971). *Procedures as a representation for data in a computer program for understanding natural language.* Massachusetts Institute of Technology, Project MAC.