

Thomas Kluth

Modeling the Contribution of Visual Attention to Spatial Language Verification

MODELING THE CONTRIBUTION
OF VISUAL ATTENTION TO SPATIAL
LANGUAGE VERIFICATION

Thomas Kluth

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doktor der Philosophie (Dr. phil.)
at the Faculty of Linguistics and
Literary Studies
Bielefeld University

May 2019

Examination Committee:

Prof. Dr. Pia Knoeferle, Humboldt University, Berlin (Reviewer)

PD Dr. Holger Schultheis, University of Bremen (Reviewer)

Prof. Dr. David Schlangen, University of Potsdam (Reviewer)

apl. Prof. Dr. Joana Cholin, Bielefeld University

Prof. Dr. Petra Wagner, Bielefeld University

Dr. Annett Jorschick, Bielefeld University

submitted on September 28, 2018

defended on May 10, 2019

© 2019 Thomas Kluth

Except where otherwise noted, this work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (<https://creativecommons.org/licenses/by-sa/4.0/>).



Gedruckt auf alterungsbeständigem Papier gemäß ISO 9706.

Pure Vernunft darf niemals siegen.
(Tocotronic, 2005)

CONTENTS

ACKNOWLEDGMENTS ix

ABSTRACT xi

ZUSAMMENFASSUNG xvii

I MOTIVATION

1	INTRODUCTION	3
1.1	Spatial Language	5
1.1.1	Spatial Prepositions	5
1.1.2	Language and Perception: The Case for Space	8
1.1.3	Spatial Prepositions and Attentional Shifts	10
1.1.4	The AVS Model	13
1.2	Thesis Outline	15
2	NON-LINGUISTIC PROCESSING OF SPATIAL RELATIONS	17
2.1	Visual Perception and Attention	17
2.1.1	Units of Visual Attention	19
2.2	Processing of Spatial Relations	22
2.2.1	Computational Framework by Logan and Sadler (1996)	23
2.2.2	Categorical and Coordinate Spatial Relations	24
2.2.3	Shifting Attention to Process Spatial Relations	26
2.3	The Type of Attention in the AVS Model	28

II COMPUTATIONAL AND EMPIRICAL STUDIES

3	THE REVERSED AVS MODEL	33
3.1	Motivating rAVS Variations	34
3.1.1	Comparison to PC(-BB) Models	38
3.2	Model Evaluation	40
3.2.1	Goodness-of-Fit and Simple Hold-Out: Method	41
3.2.2	Logan and Sadler (1996) and Hayward and Tarr (1995)	45
3.2.3	Proximal and Center-of-Mass Orientation (Exps. 1–3)	49
3.2.4	Dissociate Center-of-Mass from Midpoint (Exp. 4)	59
3.2.5	Grazing Line Effect (Exps. 5 & 6)	62
3.2.6	Effect of Distance (Exp. 7)	67
3.2.7	All Experiments from Regier and Carlson (2001)	72
3.3	Discussion of Evaluation of rAVS Variations	73
4	EMPIRICALLY ASSESSING MODEL PREDICTIONS	75
4.1	Predictions	76
4.1.1	Relative Distance	76
4.1.2	Asymmetrical ROs	78
4.1.3	Parameter Space Partitioning	81
4.2	Empirical Study	85
4.2.1	Results: Acceptability Ratings	91
4.2.2	Results: Eye Movements	104

4.2.3	Results: Reaction Times	110
4.2.4	Discussion of the Empirical Study	111
5	MODEL SIMULATIONS	115
5.1	Implementing the Preference for the Center-of-Object	116
5.1.1	The AVS-BB Model	116
5.1.2	The rAVS-CoO Model	117
5.2	Fitting Models to Data: GOF and SHO	118
5.2.1	Motivation for Global Model Analyses	120
5.3	Parameter Space Partitioning: Center-of-Object Models	121
5.4	Model Flexibility Analysis	123
5.4.1	Method	124
5.4.2	Results	128
5.5	Landscaping	131
5.5.1	Method	131
5.5.2	Results	132
5.6	Discussion of All Model Simulations	138
5.7	Outlook: Rating Distributions and Bayesian Inference	139
5.7.1	Rating Distributions	139
5.7.2	Bayesian Inference Using the Cross-Match Test	142
III GENERAL DISCUSSION		
6	TOWARDS A COMPREHENSIVE MODEL OF SPATIAL LANGUAGE PROCESSING	151
6.1	Summary of Findings	151
6.2	Levels of Analysis: Marr's Three-Level Proposal	153
6.2.1	AVS-like Models and Marr's Levels	154
6.2.2	Extending the Computational Level	155
6.2.3	Explicating the Algorithmic and Representational Level	157
6.2.4	Extending the Implementational Level	165
6.3	Summary of Ideas for Future Model Enhancements	167
6.4	Conclusion: Does Directionality of Attention Matter?	168
IV APPENDIX		
A	LIST OF ABBREVIATIONS	173
B	EMPIRICAL STUDY	175
C	MODEL FLEXIBILITY ANALYSIS	177
D	DEFENSE THESES	179
E	IMAGE CREDITS	185
F	LIST OF FIGURES	187
G	LIST OF TABLES	197
	BIBLIOGRAPHY	201

ACKNOWLEDGMENTS

Earning a Ph.D. is an adventurous journey. Enthusiastically, you start with seven-league boots. You are impressed about all the new things that lie at the wayside. Slowly, however, the road gets bumpy and it starts to rain. In the darkness of the forest you struggle with grumpy little creatures – and ask yourself why you left your cozy home in the first place. Fortunately, glittering gems of understanding illuminate your path. As you fight against your mightiest opponent, dawn begins. Finally, you struck him down, sun rises, and you successfully collect the treasure.

Adventures only end well, if numerous people support and help the protagonist. I have experienced that during the last four years of my life. I would like to express my gratitude to everyone who supported me along my way.

First and foremost, I would like to thank my supervisor Pia Knoeferle. Throughout the years she supported me in an excellent way: providing freedom to think and do what I deemed right, improving my texts by giving detailed feedback, responding super-fast to e-mails (like walking into her office despite the physical separation), and – every now and then – stimulating my intellectual journey for weeks or months with only one short sentence. I cannot imagine a better supervision.

At the CITEC, Michele Burigo supported and encouraged me a great deal. Among other things, he taught me how to conduct a psychological experiment, improved my texts for non-computational readers, convinced me to please reviewers, recommended the *Hyperion Cantos* from Dan Simmons to me, and encouraged me to carry on when everything felt pointless.

Holger Schultheis – who initialized my interest in cognitive modeling back in 2013 – continuously provided support from Bremen. His objective and honest feedback saved me from continuing to work with short-sighted thoughts (e.g., statistical nonsense or long-winded justifications based on false assumptions), his ideas improved the models, and he provided me with access to computing resources.

My colleagues in the former Language & Cognition group at the CITEC (Alba Rodríguez Llamazares, Dato Abashidze, Eva Maria Nunemann, Julia Marina Kröger, Katharina Wendler, and Katja Münster) warmly welcomed me – a computer scientist – to the realms of psycholinguistics and the life of Ph.D. students. Thank you!

I wish to thank the members of my examination committee for their time and efforts, the CITEC Graduate School for the intellectual and organizational infrastructure, and the DFG for the financial support.

Thanks also to all creators and maintainers of the numerous free and open source software tools that facilitated my research.

Finally, I want to thank my friends and my family for the continuous encouragement during the last years. It is wonderful to be surrounded by so many great people! In particular, I want to thank my fiancée Laura: Without you at my side, I would not have been able to finish this Ph.D. I am very much looking forward to continuing our shared journey!

ABSTRACT

This research asks how humans connect spatial language to physical space. To investigate this question, the present dissertation focuses on the task of verifying sentences containing a projective spatial preposition (e.g., *above*, *below*) against a depicted spatial relation (e.g., a circle above a rectangle). Linguistically, the two components of a spatial relation are distinguished from each other: “The [located object (LO)] is above the [reference object (RO)].” That is, a spatial preposition specifies the location of an LO with respect to an RO. Typically, semantics do not allow to interchange RO and LO (although syntactically this is not a problem). For instance, compare the sentence “The bike (LO) is in front of the house (RO)” with “The house (LO) is behind the bike (RO)” (cf. Talmy, 2000, p. 183).

For the processing of spatial relations, shifts of visual attention have been identified as an important mechanism (Franconeri, Scimeca, Roth, Helseth, & Kahn, 2012; Logan & Sadler, 1996; see Chapters 1 and 2). While Logan (1995) and Logan and Sadler (1996) claimed that attention should shift from the RO to the LO during the processing of spatial relations, recent empirical evidence suggests that the shift of attention might also take place in the same order as the sentence unfolds – from the LO to the RO (Burigo & Knoeferle, 2015; Roth & Franconeri, 2012).

A computational cognitive model of spatial language verification is the ‘Attentional Vector Sum’ (AVS) model proposed by Regier and Carlson (2001). This model (implicitly) implements a shift of attention from the RO to the LO (see Chapter 1). It accommodates empirical data from a range of different spatial RO-LO configurations (Regier & Carlson, 2001). To what extent does this good model performance originate from the directionality of the implemented shift (from the RO to the LO)? Considering the recent empirical evidence that attention might move in the reversed direction (from the LO to the RO) – would a model implementing such a reversed shift perform better or worse on the empirical data? These are the main questions that motivated the present thesis.

To answer these questions, I developed several variations of the AVS model (taking into account the two important geometric properties ‘proximal orientation’ and ‘center-of-mass orientation’; Regier, 1996; Regier & Carlson, 2001). In all these variations, the shift of attention goes from the LO to the RO (instead of from the RO to the LO). This is why they are called ‘reversed AVS’ (rAVS) models. In Chapter 3, I assess the rAVS variations using empirical data (acceptability ratings for spatial prepositions) from Hayward and Tarr (1995), Logan and Sadler (1996), and Regier and Carlson (2001). More specifically, I fitted the

models to the empirical data (separately for each experiment and for the whole data set from Regier & Carlson, 2001). That is, I minimized the ‘normalized Root Mean Square Error’ (nRMSE) and thus obtained a ‘goodness-of-fit’ (GOF) measure. Moreover, I evaluated the ability of the models to generalize to unseen data (cf. Pitt & Myung, 2002) by applying the ‘simple hold-out’ method (SHO; Schultheis, Singhaniya, & Chaplot, 2013). The SHO is a cross-fitting method that accounts for potential over-fitting of empirical data. Considering these model benchmarks, one rAVS variation – the $rAVS_{w-comb}$ model – performs as well as the AVS model on the tested empirical data. The $rAVS_{w-comb}$ model implements a mechanism in which ‘relative distance’ (roughly: absolute distance from LO to RO divided by the dimensions of the RO) weights the influence of the two important geometric features proximal orientation and center-of-mass orientation. Based on these results, neither implementation of directionality of attention is able to accommodate the empirical findings better than the other.

This is why I analyzed the AVS and $rAVS_{w-comb}$ models in terms of their predictions (Chapter 4). The idea was to identify stimuli for which the two contrasting shift-implementations (i.e., the two models) predict different outcomes. Data collected with these stimuli could then potentially tell apart the two models (e.g., if humans follow predictions from one model but not from the other). I created two types of test cases for which the two models seemed to generate somewhat different outcomes: a relative distance test case and an asymmetrical ROs test case.

In the relative distance test case, the critical manipulation is the height of the rectangular ROs. The absolute placements of the LOs remain equal in these stimuli. This test case is the first to investigate a potential influence of relative distance on human spatial language acceptability ratings. The predictions for the relative distance test case were that across different RO heights, acceptability ratings should differ (despite equal absolute LO placements). This prediction was clear for the $rAVS_{w-comb}$ model. However, due to the averaging vector sum mechanism in the AVS model, the prediction from the AVS model remained unclear.

The second test case (asymmetrical ROs) challenges the role of the vector sum in the AVS model. For this test case, I designed asymmetrical ROs. LOs are placed either above the cavity of these ROs or above the mass. (The RO-side that faces the LO is flat.) For these ROs, the center-of-mass does not coincide with the center-of-object (the center of the bounding box of the RO). Based on intuitive reasoning, the AVS model predicts different acceptability ratings for LOs placed (i) with equal distance to the center-of-mass but (ii) either above the cavity or the mass of the RO: the AVS model seems to predict higher ratings for LOs placed above the mass compared to LOs above the cavity. The $rAVS_{w-comb}$ model predicts no difference for this test case.

I systematically simulated the models on the created stimuli using the ‘Parameter Space Partitioning’ method (PSP; Pitt, Kim, Navarro, & Myung, 2006). This method enumerates all qualitatively different data patterns a model is able to generate – based on evaluating the whole parameter space of the model. Surprisingly, the PSP analysis revealed that both models share some of their predictions (but the models do not generate equal outcomes for all stimuli and parameter settings). Empirical data collected with these stimuli still might help to distinguish between the two models in terms of performance (e.g., based on different quantitative model fits).

This is why I conducted an empirical study that tested the model predictions for both developed test cases (relative distance and asymmetrical ROs). The empirical study was designed to be as close as possible to the experimental setup reported in Regier and Carlson (2001). That is, 34 participants read the German sentence “Der Punkt ist über dem Objekt” (“The dot is above the object”) and afterwards had to rate its acceptability given a depicted spatial relation (e.g., an image of a dot and a rectangle) on a scale from 1 to 9. In addition to *über* (*above*), I also tested the German preposition *unter* (*below*). In total, the study tested 448 RO-LO configurations. Moreover, I tracked the eye-movements of participants during inspection of the depicted spatial relation. These data are a measure of overt attention during spatial relation processing.

The empirical study could generalize effects on spatial language verification from English to German (‘grazing line’ effect and lower ratings for *unter*, *below*, compared to *über*, *above*). Furthermore, the empirical study revealed an effect of relative distance on spatial language acceptability ratings, although different than predicted by the $rAVS_{w-comb}$ model. The empirical data from the rectangular ROs suggest that lower relative distance weakens (i) the effect of proximal orientation and (ii) – for high values of proximal orientation – weakens a reversed effect of center-of-mass orientation. Neither the $rAVS_{w-comb}$ model nor the AVS model can fully accommodate this finding. Future research should more closely investigate the effect of relative distance.

For the asymmetrical ROs, analyses of the empirical data suggest that people rely on the center-of-object instead of on the center-of-mass for their acceptability ratings. This challenges earlier findings about the importance of the center-of-mass orientation. However, given that in earlier studies, the center-of-mass and the center-of-object most often coincided, the data presented in this dissertation provide additional information on how humans process geometry in the context of spatial language verification.

In terms of eye movements, the empirical data provide evidence for the horizontal component of the attentional focus as defined in the AVS model. This focus is also an important point in the $rAVS_{w-comb}$ model. The empirical results do not contradict the vertical component of the hypothesized attentional focus. However, due to the design of the study,

it remains unclear whether the vertical fixation locations were caused by the used preposition or by the vertical location of the LO. In addition, people inspected the two types of asymmetrical ROs slightly differently. For the more open asymmetrical shapes (L-shaped), fixations were influenced by the asymmetrical distribution of mass. In contrast, for the less open but still asymmetrical shapes (C-shaped), fixation patterns could not be distinguished from fixation patterns to rectangular ROs. Note that for all asymmetrical ROs, the center-of-object orientation could predict the rating data better than the center-of-mass orientation – despite distinct fixation patterns.

To further analyze the claim that people might use the center-of-object instead of the center-of-mass for their ratings, I developed modifications for the two cognitive models. While the AVS and $rAVS_{w-comb}$ models rely on the center-of-mass, the two new models ‘AVS bounding box’ (AVS-BB) and ‘ $rAVS$ center-of-object’ ($rAVS-CoO$) consider the center-of-object instead (the rest of the models remains unchanged). To thoroughly analyze all four cognitive models, I applied several model comparison techniques (Chapter 5). Based on the stimuli and data from the empirical study, the goal of the model simulations was to distinguish between models that implement a shift from the RO to the LO (AVS, AVS-BB) and models that implement a shift from the LO to the RO ($rAVS_{w-comb}$, $rAVS-CoO$). Apart from fitting the models to the data (per GOF and SHO), I analyzed them using the ‘Model Flexibility Analysis’ (MFA, Veksler, Myers, & Gluck, 2015) and the ‘landscaping’ method (Navarro, Pitt, & Myung, 2004). The latter two methods provide information on how flexible the models are. A highly flexible model is able to generate a vast amount of distinct output. A model with low flexibility generates only few distinct data patterns. In comparing model performances, one should consider the model flexibility (Roberts & Pashler, 2000). This is because a more flexible model might even fit empirically implausible data well – due to its high flexibility. This renders a close fit to empirical data a necessary but not sufficient criteria for a “good” model. In addition to providing a different perspective on model flexibility, landscaping measures to what extent two models are mimicking each other (in which case it is more difficult to distinguish between them).

Considering all model simulations, the two newly proposed models $rAVS-CoO$ and AVS-BB (accounting for the center-of-object instead of for the center-of-mass) perform substantially better than their predecessors $rAVS_{w-comb}$ and AVS. In contrast to the center-of-mass models, the two center-of-object models better fit the empirical data (GOF, SHO) while they are less flexible (MFA, landscaping) and generate rating patterns closer to the empirical patterns (PSP). This supports the hypothesis that people rely on the center-of-object orientation instead of on the center-of-mass orientation. In terms of the main research question, however, the model simulations do not favor any of the two

implemented directionalities of attention over the other. That is, based on the existing empirical data and the cognitive models, both directionalities of attention are equally likely. The thesis closes with a model extension that allows cognitive modelers to analyze the models more fine-grained in the future. More specifically, extended models generate full rating distributions instead of mean ratings. This makes it possible to use all information available in the empirical data for future model assessments.

Finally, Chapter 6 summarizes the results of this Ph.D. project. Following the seminal three-level framework proposed by Marr (1982), I discuss the findings and relate them to other relevant research. I sketch several promising possibilities to enhance the models in order to create a more comprehensive model of spatial language processing. Such a model would allow cognitive scientists to further investigate how humans ground their spatial language in the visual world.

ZUSAMMENFASSUNG

Diese Dissertation beschäftigt sich mit der Frage, wie Menschen räumliche Sprache mit der äußeren Welt in Beziehung setzen. Um diese Frage zu beantworten, habe ich untersucht, wie Menschen Sätze mit lokativen räumlichen Präpositionen (z. B. *über*) angesichts einer abgebildeten räumlichen Relation (z. B. ein Punkt über einem Rechteck) verifizieren. Die lokative räumliche Präposition ordnet den beiden Objekten der räumlichen Relation verschiedene Rollen zu: „Das [zu-lokalisierende-Objekt (LO)] ist über dem [Referenzobjekt (RO)]“. Die räumliche Präposition beschreibt also den Ort des LOs in Relation zum RO. Obwohl die Syntax es zulässt, schränkt die Semantik normalerweise das Vertauschen von RO und LO ein: Während der Satz „Das Fahrrad (LO) befindet sich vor dem Haus (RO)“ nicht unüblich ist, wirkt der Satz „Das Haus (LO) befindet sich hinter dem Fahrrad (RO)“ ungewöhnlich (vgl. Talmy, 2000, S. 183).

Wissenschaftler haben Verschiebungen von visueller Aufmerksamkeit als einen wichtigen Mechanismus zur Verarbeitung von räumlichen Relationen identifiziert (Franconeri, Scimeca, Roth, Helseth, & Kahn, 2012; Logan & Sadler, 1996; s. Kapitel 1 und 2). Die Richtung der Aufmerksamkeitsverschiebung ist allerdings umstritten. Während in älteren Arbeiten eine Aufmerksamkeitsverschiebung vom RO zum LO angenommen wurde (Logan, 1995; Logan & Sadler, 1996) haben jüngere empirische Befunde gezeigt, dass sich Aufmerksamkeit möglicherweise eher in der Reihenfolge des Satzes verschiebt – d. h. vom LO zum RO (Burigo & Knoeferle, 2015; Roth & Franconeri, 2012).

Das ‚Attentional Vector Sum‘-Modell (AVS, Aufmerksamkeitsvektorensumme, Regier & Carlson, 2001) ist ein komputationales, kognitives Modell der Verifizierung räumlicher Sprache. Dieses Modell nimmt (implizit) an, dass sich Aufmerksamkeit vom RO zum LO verschiebt (s. Kapitel 1). Das Modell kann die empirischen Daten einer Reihe von verschiedenen räumlichen RO-LO Konfigurationen gut abbilden (Regier & Carlson, 2001). Inwieweit hängt dieser Modellerfolg von der implementierten Richtung (vom RO zum LO) der Aufmerksamkeitsverschiebung ab? Wenn man die jüngsten empirischen Befunde in Betracht zieht, die stattdessen eine Aufmerksamkeitsverschiebung vom LO zum RO nahelegen: Würde ein Modell, welches eine Aufmerksamkeitsverschiebung vom LO zum RO implementiert, die empirischen Daten besser oder schlechter abbilden? Dies sind die Hauptforschungsfragen, die dieser Dissertation zu Grunde liegen.

Um diese Fragen zu beantworten, habe ich mehrere Variationen des AVS Modells entwickelt. In allen Variationen ist eine Aufmerksamkeitsverschiebung vom LO zum RO implementiert – unter Berücksichtigung

der geometrischen Faktoren ‚proximal orientation‘ und ‚center-of-mass orientation‘, von denen bekannt ist, dass sie die Akzeptanz von räumlichen Präpositionen beeinflussen (Regier, 1996; Regier & Carlson, 2001). Das Umkehren der Richtung der Aufmerksamkeitsverschiebung spiegelt sich im Namen der neuen Modellvariationen wider: Ich habe sie ‚reversed AVS‘-Modelle (rAVS, umgekehrte AVS-Modelle) genannt. In Kapitel 3 habe ich alle rAVS-Variationen daraufhin untersucht, ob sie bereits existierende empirische Daten nachbilden können (Daten von Hayward & Tarr, 1995; Logan & Sadler, 1996; Regier & Carlson, 2001). Diese Daten sind Akzeptanzbewertungen von räumlichen Präpositionen angesichts abgebildeter räumlicher Relationen.

Ich habe alle Modelle simuliert, um zu analysieren, wie gut die Modelle ihre künstlichen Daten an die empirischen Daten anpassen können (Daten von jedem Experiment einzeln sowie den gesamten Datensatz von Regier & Carlson, 2001). Das heißt, dass ich die Abweichung zwischen den empirischen und den modellgenerierten Daten minimiert habe (genauer: den ‚normalized Root Mean Square Error‘, nRMSE, also die normalisierte Wurzel aus der mittleren quadratischen Abweichung). Dies liefert eine Güte der Modellanpassung (‚goodness-of-fit‘, GOF). Darüber hinaus habe ich untersucht, wie gut die Modelle angesichts ungesehener Daten in der Lage sind, zu generalisieren (vgl. Pitt & Myung, 2002). Dazu habe ich die ‚simple hold-out‘-Methode genutzt (SHO, einfaches Weglassen; Schultheis, Singhaniya, & Chaplot, 2013). Die SHO-Methode ist eine Kreuzvalidierungsmethode, die eine mögliche Überanpassung (‚over-fitting‘) berücksichtigt. Die Modellevaluation mithilfe dieser Methoden hat gezeigt, dass eine rAVS-Variation – das rAVS_{w-comb}-Modell – die getesteten Daten genauso gut abbilden kann wie das AVS-Modell. Das rAVS_{w-comb}-Modell benutzt dazu ‚relative Distanz‘ (grob: absolute Distanz zwischen LO und RO dividiert durch die Abmessungen des ROs), um den Einfluss der beiden geometrischen Faktoren ‚center-of-mass orientation‘ und ‚proximal orientation‘ zu gewichten. Diese Ergebnisse bedeuten, dass keine der beiden Richtungen der Aufmerksamkeitsverschiebung die empirischen Daten besser erklären kann als die andere.

Deshalb habe ich die AVS- und rAVS_{w-comb}-Modelle daraufhin untersucht, ob sie eventuell unterschiedliche Datenmuster für noch nicht getestete RO-LO Konfigurationen vorhersagen (Kapitel 4). Wenn das der Fall wäre, könnten empirische Daten für diese Stimuli dabei helfen, zwischen den beiden Modellen – Implementierungen gegensätzlicher Richtungen der Aufmerksamkeitsverschiebung – zu unterscheiden (z. B. indem die Vorhersage des einen Modells aber nicht die des anderen Modells erfüllt wird). Ich habe zwei Testfälle entwickelt, für die die beiden Modelle den Anschein machten, unterschiedliche Datenmuster vorherzusagen. Der eine Testfall betrifft die relative Distanz, der zweite Testfall untersucht die Rolle von asymmetrischen RO.

Im Testfall zur relativen Distanz ist die kritische Manipulation, dass ich Rechtecke mit verschiedenen Höhen als RO genutzt habe. Die absolute Platzierung der LO bleibt konstant für alle Rechtecke. Die in dieser Arbeit präsentierte Studie ist die erste, die einen möglichen Einfluss von relativer Distanz auf Akzeptanzbewertungen von räumlichen Präpositionen untersucht. Das $rAVS_{w-comb}$ -Modell sagt klar voraus, dass sich die Akzeptanzbewertungen zwischen den verschiedenen hohen Rechtecken unterscheiden sollten (trotz gleicher absoluter Platzierung der LO). Die Vorhersage des AVS-Modells bleibt unklar. Ein Hauptgrund für diese Unklarheit ist die Vektorensomme, die über die Geometrie des ROs mittelt.

Der zweite Testfall untersucht den Einfluss von asymmetrischen RO. Hier steht insbesondere die Vektorensomme des AVS-Modells im Fokus, die dafür verantwortlich ist, die Geometrie des ROs abzubilden. Ich habe die asymmetrischen RO so entwickelt, dass LO, die über den asymmetrischen RO platziert werden, entweder über dem Hohlraum des ROs oder über Masse des ROs liegen. (Die Seite des ROs, die zum LO zeigt, ist flach.) Der Schwerpunkt des ROs („center-of-mass“) stimmt nicht mit dem Mittelpunkt des ROs („center-of-object“) überein. Der Mittelpunkt ist die Mitte des kleinsten Rechtecks, das alle Punkte des ROs beinhaltet (der sogenannten „bounding box“). Intuitiv sagt das AVS-Modell voraus, dass zwei LOs, die mit gleicher Distanz zum Schwerpunkt aber entweder über dem Hohlraum oder über der Masse des asymmetrischen ROs platziert werden, unterschiedlich bewertet werden sollten. Konkreter sagt das AVS-Modell voraus, dass das LO, welches sich über der Masse befindet, höher bewertet werden sollte als das LO, welches sich über dem Hohlraum befindet. Das $rAVS_{w-comb}$ -Modell sagt keinen Unterschied in Bewertungen für diesen Testfall voraus.

Mithilfe der „Parameter Space Partitioning“-Methode (PSP, Parameter-Raum-Aufteilung, Pitt, Kim, Navarro, & Myung, 2006) habe ich die Modelle systematisch untersucht. Diese Methode identifiziert alle vorhergesagten Datenmuster eines Modells, die sich qualitativ unterscheiden. Dazu durchsucht die PSP-Methode den gesamten Parameterraum des Modells. Überraschenderweise stellte sich durch diese Methode heraus, dass beide Modelle (AVS und $rAVS_{w-comb}$) überlappende Vorhersagen treffen. (Das heißt nicht, dass beide Modelle mit allen Parametersätzen und für alle Stimuli genau die gleichen Vorhersagen treffen.) Trotz der teilweise überlappenden Vorhersagen könnten empirische Daten für diese Stimuli dabei helfen, die beiden Modelle voneinander zu unterscheiden (z. B. durch quantitativ unterschiedliche Modellanpassungen an die Daten).

Deshalb habe ich eine empirische Studie mit diesen Stimuli durchgeführt, um die Vorhersagen der Modelle hinsichtlich der beiden vorgestellten Testfälle (relative Distanz und asymmetrische RO) zu überprüfen. Die Studie wurde so gestaltet, dass sie möglichst gut vergleichbar mit früheren Studien ist (insbesondere mit den Experimenten von Re-

gier & Carlson, 2001). 34 Studienteilnehmer sollten den Satz „Der Punkt ist über dem Objekt“ lesen und danach die Akzeptanz dieses Satzes hinsichtlich einer abgebildeten räumlichen Relation (also eines Bildes mit einem Punkt und einem Objekt) auf einer Skala von 1 bis 9 bewerten. Zusätzlich zur Präposition *über* habe ich die Präposition *unter* getestet. Die Studie beinhaltete insgesamt 448 verschiedene räumliche RO-LO Konfigurationen. Darüber hinaus habe ich die Augenbewegungen der Teilnehmer während der Präsentation der Raumrelationen aufgenommen. Diese stellen eine interessante Messgröße von offener visueller Aufmerksamkeit dar.

Die Studie generalisiert Effekte vom Englischen ins Deutsche („grazing-line“-Effekt und niedrigere Bewertungen für *unter* im Vergleich zu *über*). Für den Testfall der relativen Distanz zeigen die Ergebnisse der empirischen Studie, dass relative Distanz Akzeptanzbewertungen räumlicher Sprache beeinflusst. Dieses Ergebnis bestätigt die generelle Vorhersage des $rAVS_{w-comb}$ -Modells. Allerdings unterscheidet sich die empirisch gefundene Art und Weise des Effekts der relativen Distanz von dem konkreten Mechanismus des $rAVS_{w-comb}$ -Modells. Analysen der Daten legen nahe, dass niedrige relative Distanz (i) den Effekt der ‚proximal orientation‘ schwächt und dass niedrige relative Distanz (ii) – bei hohen Werten der ‚proximal orientation‘ – einen umgekehrten Effekt der ‚center-of-mass orientation‘ schwächt. Da weder das AVS-Modell noch das $rAVS_{w-comb}$ -Modell diesen Mechanismus erklären kann, sollte zukünftige Forschung diesen Effekt genauer untersuchen.

Für den Testfall der asymmetrischen RO legen die Daten nahe, dass Menschen statt des Schwerpunkts des ROs (‚center-of-mass‘) eher den Mittelpunkt des ROs (‚center-of-object‘) als Basis für ihre linguistischen Akzeptanzbewertungen nehmen. Dieses Ergebnis stellt die Bedeutung der ‚center-of-mass orientation‘ in Frage und lässt es wahrscheinlicher erscheinen, dass Menschen sich auf die ‚center-of-object orientation‘ stützen. Da allerdings in den meisten vorherigen Studien Schwer- und Mittelpunkt zusammenfielen, geben die hier vorgestellten Daten interessante neue Einblicke in die Art und Weise, wie Menschen asymmetrische Objekte zur Verifizierung von räumlichen Ausdrücken verarbeiten.

Die gesammelten Augenbewegungsdaten bestätigen die horizontale Komponente des im AVS-Modell definierten ‚Aufmerksamkeitsfokus‘ (dieser Punkt spielt auch im $rAVS_{w-comb}$ -Modell eine wichtige Rolle). Obwohl die Daten nicht der vertikalen Komponente dieses Fokus‘ widersprechen, lässt sich durch das Studiendesign nicht zweifelsfrei klären, ob die Präposition oder die Platzierung der LO die vertikalen Fixationen beeinflusst hat. Darüber hinaus haben die Augenbewegungsdaten gezeigt, dass die Studienteilnehmer die beiden unterschiedlichen Typen der asymmetrischen RO unterschiedlich inspiziert haben. Während die Augenbewegungen durch die asymmetrische Massenverteilung der offeneren asymmetrischen RO (L-förmig) beeinflusst wurden, haben die Studienteilnehmer die geschlosseneren asymmetrischen RO

(C-förmig) so fixiert, als wenn diese RO rechteckig wären. Trotz dieser unterschiedlichen Fixationsmuster kann die ‚center-of-object orientation‘ die empirischen Akzeptanzbewertungen besser erklären als die ‚center-of-mass orientation‘.

Um die Hypothese, dass Menschen sich zur Verifizierung von räumlichen Präpositionen eher auf den Mittel- statt auf den Schwerpunkt des ROs beziehen, näher zu untersuchen, habe ich die beiden Modelle AVS und $rAVS_{w-comb}$ leicht modifiziert. Daraus sind die neuen Modelle ‚AVS bounding box‘ (AVS-BB) und ‚rAVS center-of-object‘ (rAVS-CoO) entstanden. Anstatt den Schwerpunkt des ROs in ihren Berechnungen zu berücksichtigen (wie AVS und $rAVS_{w-comb}$), nutzen die neuen Modelle AVS-BB und rAVS-CoO den Mittelpunkt des ROs. Die übrigen Bestandteile der Modelle sind unverändert geblieben. Um alle vier Modelle gründlich zu analysieren, habe ich eine Reihe weiterer Modellsimulationen durchgeführt (Kapitel 5). Mithilfe der Daten und Stimuli der Studie aus Kapitel 4 habe ich versucht, die Modelle, die eine Aufmerksamkeitsverschiebung vom RO zum LO implementieren (AVS, AVS-BB), von den Modellen, die eine umgekehrte Aufmerksamkeitsverschiebung (vom LO zum RO, $rAVS_{w-comb}$, rAVS-CoO) implementieren, zu unterscheiden. Dazu habe ich alle Modelle an die gesammelten empirischen Daten angepasst (GOF, SHO). Darüber hinaus habe ich zwei weitere Modellanalysen durchgeführt: Die ‚Model Flexibility Analysis‘ (MFA, Modellflexibilitätsanalyse, Veksler, Myers, & Gluck, 2015) und die ‚landscaping‘-Methode (Navarro, Pitt, & Myung, 2004). Beide Methoden liefern Messgrößen, die die Flexibilität der Modelle beschreiben.

Wenn man herausfinden möchte, welches Modell einen modellierten Prozess besser beschreibt, sollte man sich nicht nur auf eine möglichst gute Anpassung der Modelle an die empirischen Daten verlassen (z. B. per GOF; Roberts & Pashler, 2000). Vielmehr ist es auch wichtig zu untersuchen, wie flexibel die Modelle sind. Ein sehr flexibles Modell kann neben den empirischen Daten auch viele weitere Datenmuster generieren, die möglicherweise empirisch nicht plausibel sind. Ein wenig flexibles Modell generiert nur eine geringe Menge an Datenmustern (im Idealfall die empirischen). Diese Überlegungen führen dazu, dass eine gute Modellanpassung an empirische Daten zwar ein notwendiges, aber kein hinreichendes Maß von Modellgüte ist. Zusätzlich zur Messung der Modellflexibilität, misst die ‚landscaping‘ Methode noch, inwieweit sich zwei Modelle nachahmen (in welchem Fall eine Unterscheidung der Modelle erschwert ist).

Über alle Modellsimulationen hinweg lässt sich feststellen, dass die Modelle, die den Mittelpunkt in ihren Berechnungen nutzen (AVS-BB und rAVS-CoO), deutlich besser abschneiden als die Ursprungsmodelle, die auf den Schwerpunkt setzen (AVS, $rAVS_{w-comb}$). Im Vergleich mit den Schwerpunktsmodellen passen sich die Mittelpunktsmodelle besser an die empirischen Daten an (GOF, SHO), sind weniger flexibel (MFA, landscaping) und generieren Datenmuster, die näher an den

empirischen Mustern liegen (PSP). Dies unterstützt die Hypothese, dass für die Verifizierung von räumlichen Präpositionen die Mittelpunktorientierung („center-of-object orientation“) wichtiger ist als die Schwerpunktorientierung („center-of-mass orientation“). Die Hauptforschungsfrage – welche Richtung der Aufmerksamkeitsverschiebung (vom RO zum LO oder vom LO zum RO) den Prozess der Verifizierung von räumlichen Präpositionen besser erklärt – lässt sich jedoch durch die Modellsimulationen nicht abschließend beantworten. Unabhängig von der implementierten Aufmerksamkeitsverschiebung lassen sich die vorliegenden Modelle anhand der existierenden Daten nicht verlässlich voneinander unterscheiden (im Sinne einer besseren Modellierung des kognitiven Prozesses). Beide Richtungen der Aufmerksamkeitsverschiebung sind gleich wahrscheinlich. Um die Modelle präziser mit empirischen Daten vergleichen zu können, stelle ich zum Schluss eine Modellerweiterung vor, die es erlaubt, dass die Modelle statt einem einzelnen Akzeptanz-Mittelwert eine komplette Verteilung von Akzeptanzbewertungen generieren können. Zukünftige Modellevaluationen können somit alle verfügbaren Informationen aus den empirischen Daten nutzen.

Die Dissertation schließt mit einer zusammenfassenden Diskussion der erreichten Ergebnisse. Basierend auf dem einflussreichen Dreiebenen-Konzept von Marr (1982) ordne ich die Befunde in weitere relevante Forschung ein. Außerdem skizziere ich einige vielversprechende Modellerweiterungen, die sich zur Entwicklung eines umfassenderen Modells von räumlicher Sprache als nützlich erweisen könnten. Solch ein Modell würde es ermöglichen, die Art und Weise, wie Menschen räumliche Sprache in der externen Welt verankern, noch präziser zu untersuchen.

Part I

MOTIVATION

INTRODUCTION

Humans live, move, and act everyday in the physical three-dimensional space. This makes referencing spatial properties of the world an important aspect of language. This type of language is called ‘spatial language’ and it has attracted much attention during the last decades (e.g., Bloom, Peterson, Nadel, & Garret, 1996; Coventry & Garrod, 2004; Landau, 2017; Landau & Jackendoff, 1993; Levelt, 1984; Levinson, 2003; Talmy, 1983). In particular, spatial language is a fruitful area for research on how language is linked to the external world as spatial language naturally describes the outer world. This grounding of language in the world seems to be quite strong, as humans even use spatial metaphors when speaking about time (e.g., “We are moving the date of our meeting forward”; Boroditsky, 2000; Moore, 2014).

It has been proposed that spatial language might be grounded in the world via a non-linguistic ‘visual attention’ mechanism (e.g., Carlson & Logan, 2005; Coventry et al., 2010; Regier & Carlson, 2001; Roth & Franconeri, 2012). Broadly speaking, visual attention is a mechanism that enables the human visual system to selectively process relevant details of the visual world (see Section 2.1 for a more fine grained introduction to visual attention). In particular, shifts of attention have been associated with the processing of spatial relations (e.g., Franconeri, Scimeca, Roth, Helseth, & Kahn, 2012; Logan & Sadler, 1996). Linguistically, spatial relations are described with spatial prepositions, such as in “The bike is *in front of* the house” (cf. Talmy, 2000, p. 183). Linguistic research on the semantics of spatial relations distinguishes the two objects in a spatial relation based on the role they play in the relation (Talmy, 2000). More precisely, in a spatial relation, a ‘located object’ (LO) is placed relative to a ‘reference object’ (RO; Logan & Sadler, 1996).¹ For instance, in “The bike is *in front of* the house”, the bike is the LO because it is located with respect to the house (the RO).

The located object (LO) is above the reference object (RO).

Given an image of a bike in front of a house, some researchers assume that people’s attention shifts from the house (the RO) to the bike (the LO) in order to verify the description (e.g., Logan & Sadler, 1996; Regier & Carlson, 2001). In contrast, empirical evidence suggests that humans shift their attention in the reversed direction – from the bike (the LO) to the house (the RO; Burigo & Knoeferle, 2015; Roth & Franconeri, 2012). The main research question for this Ph.D. project is to investigate the role of the directionality of the shift of visual attention for the verification of spatial language.

¹ There exist several other taxonomies for this distinction, e.g., ground/figure, landmark/trajector, reference/target, or relatum/locatum. The present thesis uses the RO/LO nomenclature.

The present research lies at the cross-sections of many research fields. First of all, spatial *language* naturally concerns linguistic research. In particular, this research asks how linguistic and non-linguistic processes and representations interact with each other. This makes it part of a greater psycholinguistic endeavor of investigating language use with respect to human perceptual prerequisites and the environment in which natural language occurs. This has been dubbed ‘grounding language’ (e.g., Regier & Carlson, 2001; Roy & Mukherjee, 2005; Samuelson, Smith, Perry, & Spencer, 2011) or ‘situated language processing’ (e.g., Arbib, 2017; Gorniak & Roy, 2007; Knoeferle & Guerra, 2016; Knoeferle, Pyykkönen-Klauck, & Crocker, 2016). Broadening the view to general cognitive science, this research program can be framed in terms of ‘embodied’ or ‘grounded cognition’ (e.g., Barsalou, 2008; Caligiore & Fischer, 2013; Cangelosi, 2010; Coello & Fischer, 2015; Fischer & Coello, 2015; Harnad, 1990; Pecher & Zwaan, 2005; Pezzulo et al., 2013).

In terms of non-linguistic processes, I focused on visual attention, a research topic also investigated by cognitive psychologists (for reviews see Carrasco, 2011; Kowler, 2011). In addition, research on *spatial language* is part of spatial cognition research. More specifically, I investigated projective spatial prepositions (such as *above* and *below*), a sub-class of “relational prepositions [that] describe the location of one object in relation to another” (Coventry & Garrod, 2004, p. 8). Thus, research on the processing of spatial *relations* is a relevant subfield of spatial cognition for the present research.

Methodologically, this project mainly resides in the domain of computational cognitive modeling (Sun, 2008). Generally speaking, cognitive modelers explicate (parts of) theories about cognitive processes as mathematical models, simulate these models on empirical data, and draw conclusions about cognition based on the performances of the models. Cognitive modeling is a tool of cognitive scientists since the establishment of cognitive science and remains important until today (e.g., Fum, Del Missier, & Stocco, 2007; McClelland, 2009; Shiffrin, 2010; Sun, 2009).²

The remainder of this introductory chapter provides an overview of research on spatial language processing relevant for this thesis – starting from general aspects of spatial language use and highlighting the role of (shifts of) visual attention for spatial language processing. In Section 1.1.4, I introduce the ‘Attentional Vector Sum’ (AVS) model proposed by Regier and Carlson (2001). The AVS model is a cognitive computational model that grounds spatial language verification in visual attention. To do so, it assumes a shift of attention from the RO to

² In addition, cognitive modeling has influenced real-world technical solutions such as the technology of ‘deep learning’, which is a component of many “artificial intelligence” products. This technology originates from neural networks – i.e., cognitive models developed in the so-called ‘Parallel Distributed Processing’ or ‘Connectionist’ Framework (Mayor, Gomez, Chang, & Lupyan, 2014; McClelland, Rumelhart, & PDP Research Group, 1986; Rumelhart, McClelland, & PDP Research Group, 1986).

the LO. By modifying this assumption – i.e., reversing the directionality of the shift –, the AVS model serves as basis for my own computational and empirical studies presented in Part II. In Section 1.2, the first chapter closes with an outline of the remainder of this thesis.

1.1 SPATIAL LANGUAGE

1.1.1 *Spatial Prepositions*

Spatial language consists of more than spatial prepositions but especially prepositions were studied extensively (see Coventry & Garrod, 2004; Landau, 2017, for reviews) – the present research also concerns spatial prepositions. One major outcome of research on spatial prepositions is that their use is affected by two different forces: world knowledge and geometry. The latter should be no surprise for *spatial* prepositions. However, it is an interesting finding that world knowledge affects the use of spatial prepositions, too.

WORLD KNOWLEDGE People produce different spatial prepositions dependent on the assumed functional interaction of the RO and the LO. For instance, Feist and Gentner (2003) showed that their participants more frequently used the preposition *in* than *on* when the RO was called a bowl. In contrast, they used *on* more often than *in* when the very same RO was called a plate (see also Coventry, Carmichael, & Garrod, 1994; Vandeloise, 1991).

People also comprehend spatial prepositions with respect to how the described objects typically interact in the world. For instance, in their first experiment, Carlson-Radvansky, Covey, and Lattanzi (1999) asked their participants to place pictures of objects above/below each other. Crucially, the objects were either in a typical functional relation (e.g., a toothpaste tube and a toothbrush), or they were functionally unrelated (e.g., a tube of oil paint and a toothbrush). In contrast to what one would expect if the spatial prepositions *above/below* only code for geometric properties of the scene, Carlson-Radvansky et al. (1999) found that their participants did not place the LO (e.g., the toothpaste tube) centrally above the RO (e.g., the toothbrush). Rather, the placement of the LO deviated towards the part of the RO that functionally interacted with the LO (e.g., the bristles of the toothbrush). In a second experiment, Carlson-Radvansky et al. (1999) found higher acceptability judgments for LOs located in positions that enabled functional interaction (a coin directly above the slot of a piggy bank) vs. positions that did not enable this interaction (a coin slightly to the left or right of the slot) – despite equal geometric properties of the RO (apart from the location of the slot for different piggy banks). Hörberg (2008) conducted similar experiments using Swedish prepositions and found the same empirical pattern (see also Coventry, Prat Sala, & Richards, 2001).

Visual attention is hypothesized to unite effects of world knowledge and geometry on spatial language use.

Carlson, Regier, Lopez, and Corrigan (2006) proposed a modification of the AVS model to account for acceptability judgments influenced by world knowledge. In my master thesis, I developed and tested further extensions to integrate world knowledge into the AVS model (Kluth, 2014; Kluth & Schultheis, 2014). All these model extensions are based on the assumed role of visual spatial attention for spatial language processing. In particular, Carlson et al. (2006) argue that visual spatial attention is the mechanism that reconciles geometric and functional aspects in spatial language use.

Further evidence for the importance of visual attention for world knowledge aspects of spatial language comes from Coventry et al. (2010). In one of their experiments, they tracked participants' eye movements (i.e., overt visual attention; see Section 2.1 for an introduction into visual attention research) during a spatial language acceptability rating task. The experiment was designed to gradually manipulate the strength of the functional interaction between the RO and the LO. For instance, Coventry et al. (2010) showed images of a cornflakes box above a bowl. In the "functional" condition, the (static) image depicted cornflakes falling out of the box "at such a trajectory that they would land in the container below" (Coventry et al., 2010, p. 207). In the "non-functional" condition, the trajectory of the cornflakes indicated that they would miss the bowl. Finally, in the "control" condition, no falling cornflakes were depicted. Participants were shown a sentence like "The box is *above* the bowl" (not mentioning the cornflakes) and afterwards the image. The task was to rate the acceptability of the sentence with respect to the image. During inspection of the image, Coventry et al. (2010) tracked the eye movements of their participants. Images in the functional condition were rated higher than images in the non-functional or control condition. Regarding the role of visual attention for capturing functional interaction aspects, Coventry et al. (2010) compared eye movements in functional vs. non-functional scenes. In particular, they analyzed the region where the falling objects (e.g., cornflakes) would end up. In non-functional scenes, Coventry et al. (2010) found longer dwell times and more first fixations to the miss-region outside the bowl (where the cornflakes were expected to land in non-functional scenes) compared to functional scenes.

GEOMETRY Geometric properties of both, the RO and the LO, affect the comprehension and production of spatial relations. One line of research investigated the effects of different 'reference frames' on spatial language use. Following the influential theoretical framework by Logan and Sadler (1996, p. 499), a "reference frame is a three-dimensional coordinate system" (see Section 2.2.1 for more details on the framework). Levinson (2003, in particular Chapter 2) identified three types of reference frames (see also e.g., Levelt, 1984; Levinson, 1996; Logan & Sadler, 1996; Pederson, 2003; Tenbrink & Kuhn, 2011): an absolute

reference frame, a relative reference frame, and an intrinsic reference frame. The absolute reference frame is defined with respect to environmental influences (e.g., gravity), the relative reference frame is relative to an observer describing a scene, and the intrinsic reference frame takes an oriented object as base for parsing space (e.g., a chair). If these reference frames conflict with each other, people's use of spatial language is affected.

Imagine for example, a person lying on a couch and looking at a fallen over trashcan (with its upward side pointing to the feet of the person, Carlson-Radvansky & Irwin, 1993). There are three locations of a fly around the trashcan that might be described as “*above* the trashcan”: above with respect to gravity (absolute reference frame), above with respect to the viewer reclining on the sofa (relative reference frame), or above with respect to the up-side of the trashcan (intrinsic reference frame). Using different comprehension and production tasks, Carlson-Radvansky and Irwin (1993) found that all three reference frames were used to define *above*. However, participants preferred the absolute and relative reference frames over the intrinsic reference frame (see also Carlson, 1999). For the selection of a single reference frame, Carlson Radvansky and Jiang (1998) showed that conflicting reference frames are inhibited – a mechanism also discussed in the visual attention literature. More recently, Schultheis and Carlson (2017) presented evidence suggesting that not whole reference frames but rather single parameters of reference frames (i.e., origin, direction, orientation, scale; cf. Logan & Sadler, 1996, summarized in Section 2.2.1) compete for selection.

Choice of reference frame affects use of spatial prepositions.

Carlson-Radvansky and Logan (1997) showed that conflicting reference frames affect the regions of acceptability of spatial terms (i.e., spatial templates, cf. framework by Logan & Sadler, 1996). Modeling these data, Schultheis and Carlson (2018) present a combination of the AVS model (that computes acceptability ratings of spatial terms, see Section 1.1.4) and the ‘leaky, competing accumulator’ model (Usher & McClelland, 2001) proposed for reference frame selection by Schultheis and Carlson (2017). Assessing different variations of model combinations, Schultheis and Carlson (2018) suggest that the selection of a reference frame and the computation of the acceptability of a spatial term interact with each other.

Most research on spatial language investigated the properties of the RO. In contrast, Burigo, Coventry, Cangelosi, and Lynott (2016), Burigo and Sacchi (2013), and Burigo and Schultheis (2018) focused on the role of the LO, in particular the reference frame aligned on the LO. They found that the direction of the LO affects spatial language understanding (Burigo & Sacchi, 2013; Burigo & Schultheis, 2018). Burigo et al. (2016) argue that people consider the logical property of ‘converseness’ when using spatial relations (see also Levelt, 1984). The property of converseness is fulfilled, if both statements “A is *above*

B" and its converse "B is *below* A" are true. One way to manipulate this property is by rotating the LO by 180 degrees (e.g., "dog A is behind dog B" with two dogs looking in the same direction vs. two dogs looking at each other). When converseness was violated, Burigo et al. (2016) found lower linguistic acceptability ratings compared to when converseness was not violated.

*Angular deviation
from a reference
direction predicts
acceptability ratings
for spatial
prepositions.*

Manipulating geometrical properties of the RO and testing different locations of the LO, Gapp (1995) found that people's acceptability ratings were mostly affected by the angle between the RO and the LO. More precisely, the angular deviation from a reference direction was a good predictor of the ratings. Regier (1996) proposed that the orientations of two imaginary lines are important for the applicability of spatial prepositions: the 'center-of-mass orientation' (connecting the centers-of-mass of the LO and the RO) and the 'proximal orientation' (connecting the two objects where they are closest). Both observations are considered in the AVS model (Regier & Carlson, 2001, see Section 1.1.4) and are discussed in more detail in Part II, the main part of the present thesis. Before presenting the AVS model, however, I introduce two further aspects of spatial language research. In Section 1.1.2, I review research that investigates whether linguistic and non-linguistic representations of space interact with each other. In particular, I summarize the work by Hayward and Tarr (1995), who used stimuli and an experimental task comparable to the work reported in Part II. In Section 1.1.3, I review work considering the role of shifts of attention for the processing of spatial prepositions.

1.1.2 *Language and Perception: The Case for Space*

How does language relate to the physical world? This question has attracted many researchers. The domain of space is a particular fruitful area to investigate this question, because we act everyday in a physically perceivable space and, in addition, we speak effortlessly about space. Moreover, the interaction of spatial perception and spatial language is important for children's development. It has been shown that spatial language enhances children's spatial skills (e.g., Dessalegn & Landau, 2008, 2013; Farran & O'Leary, 2016; Gentner, Özyürek, Gürcanli, & Goldin-Meadow, 2013; Loewenstein & Gentner, 2005; Miller, Patterson, & Simmering, 2016; Miller, Vlach, & Simmering, 2017). Vice versa, it has been shown that spatial perception helps children to learn language (e.g., Carlson, 2007; Samuelson et al., 2011; Shusterman & Li, 2016; Smith, Maouene, & Hidaka, 2007).

In research with adults, non-linguistic spatial processing is affected by so-called 'image schemas' of verbs. Image schemas are graphical depictions of the meaning of verbs using abstract icons. Among other things, image schemas contain direction and orientation relating the agent with the patient of the verb (Richardson, Spivey, Edelman, &

Naples, 2001). For example, the verb *push* would be depicted with a horizontal image schema while the verb *respect* rather has a vertical image schema. Using a visual discrimination task (similar to Posner's cueing paradigm, an influential experimental paradigm in the research on visual attention, see Section 2.1) and a picture memory task, Richardson, Spivey, Barsalou, and McRae (2003) provide evidence that the orientation of the image schema of both concrete and abstract verbs affects non-linguistic spatial processing. In related research with nouns (e.g., Dudschig, Souman, Lachmair, de la Vega, & Kaup, 2013; Dunn, Kamide, & Scheepers, 2014), eye movements were facilitated when their direction was congruent with typically associated spatial locations of previously presented nouns (e.g., *sun*: up, *worm*: down). In summary, a more detailed sub-question of the language-world relation is whether humans' spatial processing abilities are based on shared representations for both linguistic and non-linguistic tasks.

To investigate this question, Hayward and Tarr (1995) conducted a series of experiments. Their first two experiments focused on linguistic categorization of space. In the first experiment, participants should freely describe depicted spatial relations. The two-dimensional spatial relations consisted of an RO in the center of a display and an LO placed at 48 different locations around the RO. For each RO-LO pair, participants were instructed to formulate a sentence that best described the spatial relation of the LO to the RO. The sentence should contain one or more spatial prepositions. However, they should "avoid using compass directions, a clock face, or the degree of angle" (Hayward & Tarr, 1995, p. 50). Hayward and Tarr (1995) found that participants most often used vertical prepositions (such as *above* and *below*) when the LO lay on the vertical axis from the RO (i.e., directly above or below the RO). Similarly, participants used horizontal prepositions for LOs on the horizontal axis of the RO. The use of vertical/horizontal prepositions declined for LOs that were not directly located on the vertical/horizontal axes (respectively).

In their second experiment, Hayward and Tarr (1995) asked different participants to rate the acceptability of the four prepositions that were most used in their first experiment (*above*, *below*, *left*, *right*) – using the same stimuli. The acceptability judgments confirmed the general pattern from the first experiment: LOs located on axes corresponding to the to-be-rated preposition were rated higher than LOs placed at other locations. The farther away the LOs were placed from the respective axes, the lower became the ratings. In addition, Hayward and Tarr (1995) found that distance from the LO to the RO affected ratings, which they interpreted as angular effects: "[The observed] pattern [...] suggests that the appropriateness of a given spatial term to a perceived spatial relationship is determined in part by the angle between the reference object and figure object [LO]." (Hayward & Tarr, 1995, p. 58, see also Gapp, 1995; Regier, 1996; Regier & Carlson, 2001). In Section 3.2.2,

I present an evaluation of computational models using the *above* data and the stimuli from the second experiment reported by Hayward and Tarr (1995).

The third and fourth experiment reported by Hayward and Tarr (1995) aimed at analyzing the non-linguistic spatial processing of the same stimuli. To this end, participants had to remember the depicted spatial relation in the third experiment. After a short interval with a distractor task, they should replicate the location of the LO based on the location of the RO. Again, Hayward and Tarr (1995) found that the axes of the RO affected participants' behavior: The closer the LO was located to the axes, the lower were the errors participants made when replicating the locations from their memory. The final experiment confirmed this finding with yet another experimental task. This time, participants had to detect whether a depicted spatial relation changed from the first brief display to the second brief display. A changed relation consisted of a small change of the location of the LO. Again, participants' performance was better, the closer the LO was placed to the axes of the RO.

In discussing their experiments, Hayward and Tarr (1995) argue that the axes of the RO define linguistic *prototypes* for spatial relations. Moreover, these prototypes also underlie non-linguistic categorization of space, explaining the enhanced performance close to the axes. In addition to these prototypical relations, Hayward and Tarr (1995) propose that both linguistic and non-linguistic relations encode qualitative *and* quantitative aspects. They justify this thought with the graded response pattern in all their experiments: While the axes of the RO clearly stood out in all tasks, the distance to the axes affected behavior in a quantitative way. Interestingly, to support their view they already point to early work in the neurological distinction of categorical vs. coordinate relation processing (namely Kosslyn et al., 1989). I summarize this neurological distinction and its relation to linguistic processing of spatial relations in Section 2.2.2. The important role of the axes of the RO (or: reference frames) for linguistic and non-linguistic processing of space is further addressed in Section 6.2.3. For now, let us focus on the main research topic of this thesis: the role of shifts of attention for spatial language processing.

1.1.3 *Spatial Prepositions and Attentional Shifts*

It has been shown that attention is necessary to process spatial relations (e.g., Franconeri et al., 2012; Logan, 1994) and that attention and spatial language are closely related with each other (e.g., Conder et al., 2017; Coventry et al., 2010; Roth & Franconeri, 2012, see Carlson & Logan, 2005, for a review). In particular, spatial relation processing has been associated with *shifts* of attention (see Section 2.2.3 for more details). Relating such serial movements of attention to the linguistic distinction

between an RO and an LO, Gordon Logan's influential research claimed that "the viewer's attention should move *from* the reference object *to* the located object" (Logan & Sadler, 1996, p. 499, emphasis in the original).³

This claim has certainly affected the direction of attention as modeled in the AVS model (from the RO to the LO, see Section 1.1.4) and it holds true for specific task demands (e.g., research in the conceptual cueing paradigm from Gibson and colleagues, where participants shift their attention from a central cue, an RO, to a peripheral target, an LO, see Section 2.2.3). However, other empirical studies challenge the importance of the *directionality* of the attentional shift (Burigo & Knoeferle, 2015; Coventry et al., 2010; Roth & Franconeri, 2012; see also Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002).

Coventry et al. (2010) combined eye-tracking with a sentence verification study (see Section 1.1.1 for a more detailed study description). Although Coventry et al. (2010) support Logan's general claim ("This is not to deny the importance of attention allocation from a RO to a LO", p. 211), they found somewhat contrary evidence: For superior prepositions (e.g., "The box is *over/above* the bowl"), most first fixations landed on the top object (the box) – the LO in the sentence. This gaze pattern suggests that people shifted their attention from the LO to the RO. However, Coventry et al. (2010) did not track eye movements during the comprehension of the sentence (sentence was presented before the visual scene). Thus, based on these data, one cannot time-lock the eye movements to the unfolding interpretation of the spatial sentence or to the processing of the spatial preposition.

Burigo and Knoeferle (2015) measured their participants' eye movements at the same time as the participants both saw a visual spatial relation and listened to a spatial description (a study in the psycholinguistic 'visual world paradigm', see e.g., Huettig, Rommers, & Meyer, 2011; Knoeferle et al., 2016, for reviews of this paradigm). In line with other research using the visual world paradigm, Burigo and Knoeferle (2015) provide evidence that people look at objects as they are mentioned. For sentences with spatial relations ("The LO is *above* the RO"), this means that first the LO should be inspected more than the RO – followed by more inspections to the RO than the LO. Indeed, this is a pattern found by Burigo and Knoeferle (2015) suggesting that people shift their overt attention from the LO to the RO. In addition, Burigo and Knoeferle (2015) showed that a shift from the RO to the LO also matters for the verification of the spatial utterances.

The directionality of a linguistically triggered attentional shift from the LO to the RO is further supported by research conducted within

Claim about directionality of attentional shift from RO to LO is challenged by recent empirical evidence suggesting a shift from LO to RO.

³ See also Logan (1995, p. 115): "The linguistic distinction between located and reference objects specifies a direction for attention to move—from the reference object to the located object."; Logan and Zbrodoff (1999, p. 72): "Implicit in this constraint on how we speak [about spatial relations] is the idea that attention goes first to the reference object and then to the located object. Thus, the contrast between located object and reference object provides direction to movements of attention (Logan, 1995)."

the theoretical shift-account proposed by Franconeri et al. (2012, introduced in detail in Section 2.2.3). Roth and Franconeri (2012) displayed spatial relations (two colored circles) and asked their participants to verify statements (e.g., “Is red *left of* green?”) as quickly as possible. The statements appeared before the display of the spatial relations. Crucially, Roth and Franconeri (2012) manipulated the covert allocation of attention by presenting one of the two objects slightly before the other object (0–233 ms). Roth and Franconeri (2012) found that people were quicker to verify the spatial relation, if the order of visual appearance matched the order of the to-be-verified description. That is, given “Is red *left of* green?”, people were faster to answer when the red circle appeared before the green circle (LO first, RO second) than when the green circle was displayed before the red circle (RO first, LO second). Again, this challenges Logan’s claim about an attentional movement from the RO to the LO and instead suggests that a reversed shift from the LO to the RO might be more plausible.

This idea matches findings from Huttenlocher and Strauss (1968). In their experiments, children had to place colored blocks according to the instructions from the experimenters. If the to-be-moved block was mentioned first (as an LO) in the instruction, children were faster and more accurate in placing the block compared to when the to-be-moved block was mentioned second (as an RO; see also Landau & Jackendoff, 1993, p. 225, for related studies with adults). This suggests that if a spatial task violates the linguistic order (RO first, LO second), it is more difficult than a task where the linguistic order is not violated (LO first, RO second).

As final support for an ordered sequence of attending the LO first and the RO second (i.e., in line with the order of mentioning), I point to computational models of spatial language use that are applied in robotic research – among other fields. Many of these models start with selecting the LO prior to the RO (e.g., Lipinski, Schneegans, Sandamirskaya, Spencer, & Schöner, 2012; Richter, Lins, Schneegans, Sandamirskaya, & Schöner, 2014; Richter, Lins, & Schöner, 2016, 2017; Roy & Mukherjee, 2005). However, note that this order is not necessarily a strict requirement for the functioning of the models.

After having reviewed this converging evidence for an attentional shift from the LO to the RO, I finally introduce the AVS model which implements a shift from the RO to the LO. In Chapter 3, I then introduce and assess a modification of the AVS model – the ‘reversed AVS’ (rAVS) model – that implements a reversed shift of attention from the LO to the RO – motivated by the experiments summarized in the present section.

1.1.4 The AVS Model

The ‘Attentional Vector Sum’ (AVS) model was proposed by Regier and Carlson (2001). It takes as input the locations of two “labeled”⁴ two-dimensional objects (RO and LO), the shape of the RO, and a spatial preposition (see Figure 1.1 for a visualization of the input as well as schematized model mechanisms). As output, the model computes an acceptability rating that represents how well the preposition describes the given relation (cf. empirical studies from Hayward & Tarr, 1995; Logan & Sadler, 1996).

The AVS model can be seen as consisting of two components: The height component and the angular component. The height component computes a value between 0 and 1 as a function of the vertical location of the LO relative to the top of the RO (for *above*⁵). “Intuitively, the top is the set of landmark points that are exposed from above: the ones that would get wet in the rain.” (Regier & Carlson, 2001, p. 274). The outcome of the height component is depicted in Figure 1.1a and formally computed as follows:

$$\text{height}(y) = \frac{\text{sig}(y - \text{hightop}_y, \text{highgain}) + \text{sig}(y - \text{lowtop}_y, 1)}{2} \quad (1.1)$$

The variable hightop_y denotes the y -coordinate of the highest point that is part of the top of the RO, the variable lowtop_y denotes the y -coordinate of the lowest point that is part of the top of the RO. The variable highgain is a free parameter of the model (to be adjusted for fitting the model to data). Finally, the sigmoid function $\text{sig}()$ is defined as:

$$\text{sig}(x, \text{gain}) = \frac{1}{1 + \exp(-x \cdot \text{gain})} \quad (1.2)$$

The angular component returns an acceptability rating as a function of angular deviation from a reference direction (see Figure 1.1b for a visualization). Let us first consider how the angular deviation is computed. To do so, the AVS model defines an attentional distribution (the shaded circular area in Figure 1.1b). This distribution consists of a specific amount of attention for every point i of the RO:

$$a_i = \exp\left(\frac{-d_i}{\lambda \cdot \sigma}\right) \quad (1.3)$$

Here, d_i is the Euclidean distance of the RO point i to the attentional focus point F , λ is another free model parameter, and σ is the Euclidean

4 The model does not decide which object is the RO and which is the LO; this is part of the input.

5 For different prepositions, the respective edge of the RO matters; additionally, for horizontal prepositions, the horizontal instead of the vertical location of the LO is considered.

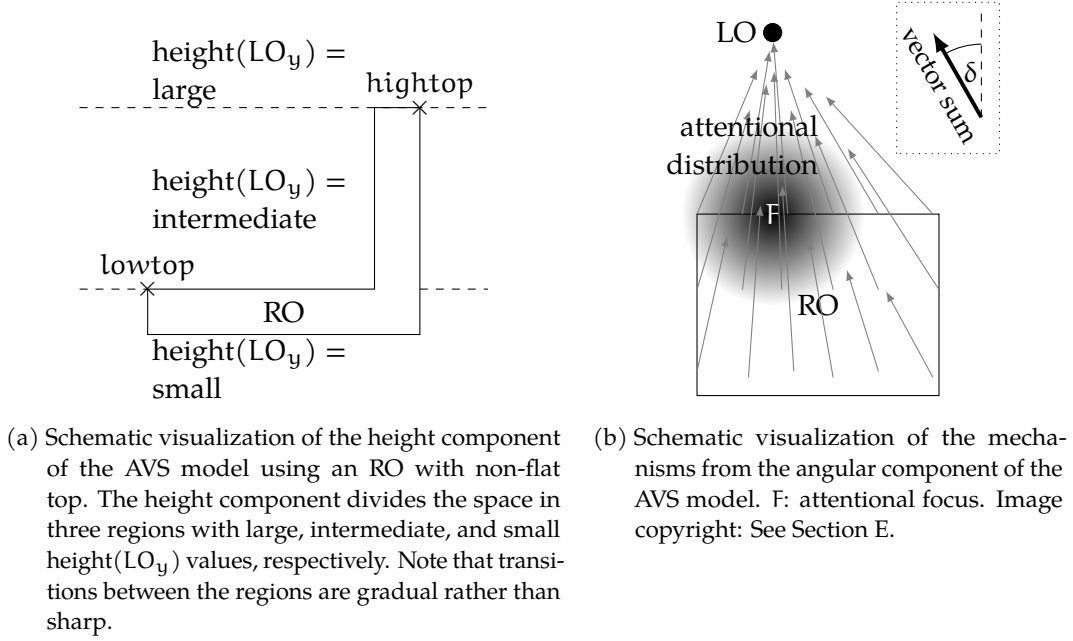


Figure 1.1: Schematic representations of (a) the height component and (b) the angular component of the AVS model.

distance between the attentional focus point F and the LO . The attentional focus point F is defined to lie on top of the RO and vertically aligned with the LO (for *above*; for different prepositions the corresponding edges of the RO are used). If the LO does not lie in the direct region above the RO , the attentional focus point is defined to be on the top-left or top-right point of the RO , respectively.

The such defined attentional distribution is used to weight a population of vectors. At every point i of the RO , a single vector $\vec{v}_i = i, \vec{LO}$ is rooted, pointing to the (point-like) LO . Every single vector v_i is multiplied with the amount of attention a_i from the attentional distribution. All vectors are summed to obtain one final vector direction (see vector labeled “vector sum” in Figure 1.1b). Formally, this process reads:

$$\overrightarrow{\text{vectorSum}} = \sum_{i \in RO} a_i \cdot \overrightarrow{i, LO} \quad (1.4)$$

To obtain an angle from this direction, the vector sum direction is compared to a reference direction (canonical upwards in the case of *above*):

$$\delta = \angle \left(\text{up}, \overrightarrow{\text{vectorSum}} \right) \quad (1.5)$$

The resulting angular deviation is used as input for a linear function that returns an acceptability rating:

$$g(\delta) = \text{slope} \cdot \delta + \text{intercept} \quad (1.6)$$

The variables slope and intercept are free model parameters. The lower the deviation, the higher the rating and vice versa. The rating from the angular component is multiplied with the outcome from the height component to obtain the final rating. The following equation describes the whole AVS model:

$$\text{above}(\text{LO}, \text{RO}) = g \left(\angle \left(\text{up}, \sum_{i \in \text{RO}} a_i \cdot \overrightarrow{i, \text{LO}} \right) \right) \cdot \text{height}(\text{LO}_y) \quad (1.7)$$

1.2 THESIS OUTLINE

The remainder of this thesis is organized as follows: In Chapter 2, further research on the processing of spatial relations is reviewed – primarily concerning non-linguistic processing. To this end, Chapter 2 starts with a summary of research on visual perception and attention (Section 2.1). Thereafter, Section 2.2.1 introduces the seminal framework on spatial relation processing by Logan and Sadler (1996). This is followed by a review of the neurological distinction of categorical and coordinate spatial relations in Section 2.2.2, highlighting the role of attention in that research. In Section 2.2.3, I summarize further evidence for the general importance of shifts of attention for spatial relation processing. The background chapter closes by applying the introduced concepts and paradigms to the conceptualization of attention in the AVS model (Section 2.3) – effectively claiming that the AVS model implements an attentional shift from the RO to the LO.

Part II contains the computational and empirical studies that I conducted within this Ph.D. project. More specifically, Section 3.1 presents several possibilities of reversing the shift of attention in the AVS model to reflect recent empirical evidence. That is, while the AVS model assumes a shift of attention from the RO to the LO, the presented alternative models implement a shift of attention from the LO to the RO. In Section 3.2, these model variations (the reversed AVS, rAVS, models) are evaluated using data from some of the literature reviewed in Chapters 1 and 2. Although the existing empirical data are sufficient to distill one rAVS variation as winner (as the others cannot accommodate all of the empirical effects), they are not sufficient for distinguishing between the winning rAVS model and the AVS model. The latter, however, concerns the main research question of this project: Does a shift from the LO to the RO (rAVS model) account better or worse for spatial language verification than a shift from the RO to the LO (AVS model)?

This is why Chapter 4 introduces an empirical acceptability rating study designed to distinguish between the two models. The stimuli for this study are based on the implications of the contrasting directionalities of the attentional shift as implemented in the two models (Section 4.1). The study reveals two novel effects on spatial language verification (relative distance and center-of-object orientation, Section 4.2). Chapter 5 presents further model modifications addressing these new effects (Section 5.1) as well as thorough computational analyses (per a variety of methods) of all competing models using the stimuli of, and the data collected in, the empirical study (Sections 5.2–5.6). In Section 5.7, the main part closes with introducing an extension that enables the models to simulate rating distributions instead of mean ratings. This model extension makes it possible to analyze the models more fine-grained in the future.

The main results of this project and their implications are discussed in Chapter 6. In addition, ideas for further research – especially model refinements – are presented.

NON-LINGUISTIC PROCESSING OF SPATIAL RELATIONS

In this chapter, I review research on non-linguistic processing of spatial relations. To this end, Section 2.1 introduces relevant work in the more general field of visual perception and attention. In particular, I highlight different conceptualizations of visual attention and the “units” attention is operating on (Section 2.1.1). These conceptualizations of attention are reconsidered in Section 2.3, in which the attentional distribution and the vector sum of the AVS model are discussed in terms of the “units of attention” they resemble.

In addition to relating the AVS model to conceptualizations of attention, Section 2.3 interprets the AVS model in the context of research explicitly asking how *shifts* of attention contribute to the processing of spatial relations. These studies are summarized in Section 2.2.

2.1 VISUAL PERCEPTION AND ATTENTION

Every day, the human visual system is accomplishing extra-ordinary work. Seemingly without efforts and time-delay, humans make sense of their visual environment, e.g., by recognizing and localizing objects. While this seems like a trivial observation, it becomes fascinating considering the vast amount of visual information processed by the retina (ca. 10^{10} bits per second, more than 1 gigabyte, Anderson, Van Essen, & Olshausen, 2005; Raichle, 2010). From this information, the visual system must quickly select the currently most relevant information given its limited resources (the visual cortex processes ca. 10^4 bits = 1.25 kilobytes per second, less than one percent compared to the retina). This selection process has been called ‘visual attention’ and it is the topic of decades of vision research (for a recent review see Carrasco, 2011).

In the following, I introduce two influential experimental paradigms in research on visual attention: ‘visual search’ and ‘spatial cueing’. Logan (1994) used the visual search paradigm to establish “that apprehending spatial relations requires spatial attention” (p. 1015). The spatial cueing paradigm was applied to investigate *shifts* of attention – from a cue to a target. In particular, using a variant of this paradigm, Logan (1995) asked how linguistic cues such as spatial prepositions control attentional shifts. Gibson and Sztybel (2014) explicitly compare linguistic with non-linguistic cues. These studies are summarized in Section 2.2.3.

Researchers used visual search and spatial cueing to investigate the role of attention for spatial relation processing.

VISUAL SEARCH A prominent experimental paradigm to investigate visual attention is the visual search paradigm (for review see Nakayama & Martini, 2011; for a comparison to the related psycholinguistic visual world paradigm see Hartsuiker, Huettig, & Olivers, 2011; Huettig, Olivers, & Hartsuiker, 2011). In the visual search paradigm, participants are given a display with several objects and a definition of a target object (e.g., its color or orientation). They have to decide as quickly as possible whether the display contains the target object or not. Performance in a visual search task is quantified via reaction time with respect to the total number of objects in the set.

Researchers found that if the target object differs only by “simple” single features from the distractor objects (e.g., a red line among green lines or a horizontal line among vertical lines), the performance does not depend on the total number of objects – the target pops-out. In contrast, if the target object is defined as a conjunction of features (e.g., a red vertical line among red horizontal and green vertical lines), reaction time increases with the number of objects in the display. It has been argued that this type of visual search (‘conjunction search’) requires humans to serially process every single item with focused attention whereas in the ‘pop-out’ or ‘feature search’, the visual scene is processed in parallel and pre-attentively (Treisman & Gelade, 1980).

Identifying spatial relations in visual search tasks resembles slow conjunction searches.

Logan (1994) used the visual search paradigm to investigate the visual processing of spatial relations. The targets were defined by their spatial relation. For instance, the target was a dash above a plus and the distractors were pluses above dashes. Logan (1994) found that using spatial relations as the defining target feature resulted in human performance similar to slow conjunction searches. Accordingly, he proposed “that apprehending spatial relations requires spatial attention” (Logan, 1994, p. 1015).

An important distinction in visual spatial attention must be made between *overt* and *covert visual attention*. Overt attention describes selecting visual features by moving the eyes while covert attention describes selection without eye movements. “Many studies have investigated the interaction of overt and covert attention, and the order in which they are deployed. The consensus is that covert attention precedes eye movements, and that although the effects of covert and overt attention on perception are often similar, this is not always the case” (Carrasco, 2011, p. 1487).

SPATIAL CUEING PARADIGM A well known experimental paradigm to investigate shifts of *covert* attention is the so-called spatial cueing paradigm (Posner, 1980; Posner, Snyder, & Davidson, 1980). In this paradigm, participants fixate a central fixation cross and are not allowed to move their eyes. Then, a cue informs about the location of a visual target that has to be recognized as quickly as possible. The cue is either central (e.g., an arrow appearing near the fixation cross pointing to the

target location) or peripheral (e.g., a dot flashing at the position where the target will appear). Central cues are said to evoke goal-directed (or endogenous, top-down) shifts of attention whereas peripheral cues are said to evoke stimulus-driven (or exogenous, bottom-up) shifts of attention (Carrasco, 2011; Corbetta & Shulman, 2002).

In the spatial cueing paradigm, the cues are valid in a majority of the trials – but not always. This is done to create trials where the target appears in a non-cued (presumably non-attended) location. In addition to central and peripheral cues, neutral cues (i.e., trials without location information) are tested – to constitute a baseline with undirected attention against which the reaction time benefits of attended locations are compared. Empirical results from the spatial cueing paradigm show that people are faster to identify the target with a valid cue and slower in trials with invalid cues – compared to trials with neutral cues, respectively.

In interpreting their results, Posner et al. (1980) likened spatial attention to a ‘spotlight’. That is, attention is seen as enhancing processing in specific spatial areas (compared to un-attended areas) just like spotlights brighten specific areas of an ongoing theater play. A related metaphor of attention¹ is the gradient model of attention (e.g., Downing & Pinker, 1985; LaBerge & Brown, 1989). Different from the spotlight model, in which the attentional strength is homogeneously distributed in the whole “brightened” area, the gradient model proposes that the attentional strength gradually decreases from its central point, the attentional focus. It is this conceptualization of attention that Regier and Carlson (2001) implemented in their AVS model (see Section 2.3).

In a paradigm similar to Posner’s spatial cueing paradigm, Logan (1995) explicitly investigated how linguistic cues control attentional shifts (see also Logan, 1994, expts. 3 and 4). More recently, Gibson and colleagues used the same paradigm to continue this line of research (for review see Gibson & Sztybel, 2014). I present these studies in more detail in Section 2.2.3.

The spatial cueing paradigm gave rise to the attention-as-a-gradient metaphor.

2.1.1 Units of Visual Attention

So far, I presented attention as filtering relevant information from the vast amount of available perceptual input (e.g., in the visual search paradigm) or as enhancing perception in specific spatial areas (e.g., in the spatial cueing paradigm). In the visual search paradigm, attention selects *objects* based on *features*. In contrast, the spatial cueing paradigm shows that attention can be directed to *spatial locations*. This brings up the question what attention selects or operates on. In other words, what are the “units” of visual attention? The answers to this question inform

¹ See Fernandez-Duque and Johnson (1999, 2002), for insightful discussions about the role of metaphors in research on attention.

the interpretation of the AVS model in terms of its representation of attention (see Section 2.3).

There is evidence that attention operates on all three dimensions italicized above. Researchers distinguish between feature-, space-, and object-based attention (Carrasco, 2011; Yantis, 2000). Feature-based attention describes the ability of the visual system to selectively attend to visual features (e.g., color or orientation) regardless of their location. A typical example is the feature search condition in the visual search paradigm. Space-based (or spatial) attention is described via metaphors like a spotlight or a gradient. It enhances visual processing for specific locations, as revealed through the spatial cueing paradigm.

Object-based attention selects objects rather than locations. For example, in a seminal study, Egly, Driver, and Rafal (1994) showed a display with two vertical rectangles. Participants had to respond to a target appearing at one of the four rectangle ends (i.e., left top, right top, right bottom, or left bottom). One rectangle end was visually cued beforehand. If the target appeared at the cued end, participants were faster compared to a target appearing at an uncued location. More interestingly, participants were faster to detect a target that appeared at the same object (e.g., bottom of left rectangle when top of left rectangle was cued) than a target appearing at the other object (e.g., top of right rectangle). Crucially, the spatial distance of the two uncued locations was exactly the same. Thus, these findings cannot be explained by a purely space-based account of visual attention.

Nuthmann and Henderson (2010) assessed whether fixation patterns to photographs of natural scenes could be better explained by the location of natural objects or by ‘saliency’ (a bottom-up stimulus-based approach used in computer vision models successfully predicting fixation locations). They found that people preferably fixate the center of natural objects instead of the center of proto-objects based on saliency computations. This is why they “suggest that saccade targeting and, by inference, attentional selection in scenes is object-based” (Nuthmann & Henderson, 2010, p. 1).

The finding that people preferably fixate the center of objects is in line with research on saccadic and perceptual localization (investigating how the visual system computes reference points on objects). More specifically, researchers found that the center-of-mass of objects seems to be a preferred saccadic end point and suggest ‘spatial pooling’ as a plausible relevant mechanism (e.g., Melcher & Kowler, 1999; Vishwanath & Kowler, 2003). The vector sum mechanism in the AVS model can be interpreted as a spatial pooling approach.

A model that combines space-based and object-based attention is the COntour DEtection Theory of Visual Attention (CODE TVA or CTVA) model (Bundesen, 1998; Logan, 1996; Logan & Bundesen, 1996). It is a combination of a theory of perceptual grouping by proximity (the COntour DEtection model, van Oeffelen & Vos, 1982, 1983) with a

Spatial pooling (such as the vector sum in the AVS model) is an important mechanism for object perception.

biased-competition account of visual attention (Theory of Visual Attention, TVA, Bundesen, 1990; Bundesen, Vangkilde, & Petersen, 2015) and accounts for a wide range of findings (Logan & Bundesen, 1996). However, as many other attentional theories and models, CTVA does not account for relations between objects. The reason for mentioning the CTVA here is that it is a promising candidate to more firmly connect the AVS model with theories of visual attention (discussed in Section 6.2.3). Among other things, this is because the CTVA refers to the influential framework of spatial relation processing proposed by Logan and Sadler (1996). Before presenting this framework in detail in Section 2.2.1, I introduce another “unit of attention” – ‘spatial indices’ – that is referred to by Logan and Sadler (1996).

SPATIAL INDICES Spatial indices have been proposed by several different authors (e.g., Ballard, Hayhoe, Pook, & Rao, 1997; O’Regan, 1992; O’Regan & Noë, 2001; Pylyshyn, 1989; Spivey, Richardson, & Fitneva, 2004; Ullman, 1984), although not all of them refer to them literally as ‘spatial index’ and not all proposals are fully compatible with each other (but the general idea is the same). I will refer to the spatial indexing theory proposed by Pylyshyn (1989, see also Pylyshyn, 1994, 2000, 2001, 2009) as this is the one cited by Logan and Sadler (1996) in their framework of spatial relation processing (see Section 2.2.1).

First empirical support of Pylyshyn’s theory comes from the multiple-object tracking paradigm (Pylyshyn & Storm, 1988) which serves here as an intuitive example of what spatial indices are supposed to be. In the multiple-object tracking paradigm, participants are presented with a display consisting of several simple, identical-looking objects (e.g., circles). A subset of objects is cued, e.g., by flashing them. Afterwards, all objects move randomly on the display. During the movement, the continuous path of a single object is the only property that serves as an identifying feature. That is, to track a single object one has to follow its movement. After some time, one object is highlighted again and the participant has to respond whether this object belongs to the previously cued objects or not. Pylyshyn and Storm (1988) found that participants could reliably track up to four objects.

As an explicit test against space-based accounts of visual attention, Pylyshyn (1994, p. 363) notes that participants’ “ability to track these targets and detect changes occurring on them does not generalize to non-targets or to items lying inside the convex polygon that they form (so that a zoom lens of attention does not fit the data).” Instead, he proposes that the visual system indexes the to-be-tracked objects. Thus, in the multiple-object paradigm, the visual system is hypothesized to track the spatial indices (referring to the objects) instead of the objects themselves.

In a nutshell, Pylyshyn’s indexing theory establishes spatial indices (or, in his original term: FINgers of INStantiations, FINSTs, Pylyshyn,

1989, p. 69) as units of attention. A spatial index “points to” an object in the visual world without encoding the properties of the object. Spatial indices are deployed pre-attentively, pre-conceptually, and automatic. Attention is then operating on these indices. If, for instance, the color of an indexed object should be retrieved, attention selects the corresponding index and the visual system is able to access the object and “query” its color. Based on empirical findings, the visual system can deploy around four to five spatial indices at one point in time (Pylyshyn, 2000).

Pylyshyn (2000, 2001) explicitly describes his spatial indexing theory as a necessary component of “situated vision”. He argues that “[i]ndexing visual objects is [...] the primary means for grounding visual concepts” (Pylyshyn, 2001, p. 127). More specifically, he proposes spatial indices as necessary “direct, preconceptual connection[s] between objects in the visual world (visual objects or proto-objects) and their representations in the visual system” (Pylyshyn, 2000, p. 197). This point of view makes spatial indexing a relevant theory in the research on how linguistic entities are grounded in the visual world (see also Spivey et al., 2004). Moreover, Pylyshyn (1989, p. 70) writes: “Being able to index particular features is especially important when encoding relational properties involving several places.” Accordingly, I review research on the processing of spatial relations in the following section.

*Spatial indices
ground mental
representations in the
visual world.*

2.2 PROCESSING OF SPATIAL RELATIONS

The processing of *relations* between objects is fundamental for human cognition. According to Gentner (2003, p. 196), “relational concepts are critical to higher-order cognition” and are one of the reasons “why we’re so smart” (title of Gentner, 2003). While Gentner’s research highlights the role of abstract relations (e.g., for making analogies, e.g., Gentner, 1983), other researchers also identified the importance of concrete relations such as visual spatial relations. For instance, Ullman (1984, p. 99) writes: “Spatial relations in three-dimensional space [...] play an important role in visual perception”; Hayward and Tarr (1995, p. 40) note that “spatial relations are a basic (and essential) element of several theories of object representation”; and Logan and Sadler (1996, p. 493) start their article with: “Spatial relations are important in many areas of cognitive science and cognitive neuroscience, including linguistics, philosophy, anthropology, and psychology.”

Given this importance, Logan and Sadler (1996) propose a computational framework of visual spatial relation processing. This framework continues to be important for research on spatial relations as diverse as empirical and computational investigations of the interaction between the proposed processes (Schultheis & Carlson, 2018), computational modeling using a neuronally plausible framework (Richter et al., 2017), control of spatial attention (Gibson & Sztybel, 2014), or transformations

for geographic information systems (Scheider, Hahn, Weiser, & Kuhn, 2018).

2.2.1 *Computational Framework by Logan and Sadler (1996)*

Motivated by linguistic research on the semantics of spatial relations, Logan and Sadler (1996) distinguish the two objects in a spatial relation based on the role they play in the relation: a located object (LO) is related to a reference object (RO, see Chapter 1). The computational framework by Logan and Sadler (1996) assumes *processes* operating on and with *representations*. Different “ordered combinations of representations and processes are interpreted as *programs* or *routines*” (Logan & Sadler, 1996, p. 501, emphasis in the original). This allows for a flexible use of the different components of the framework, e.g., to describe spatial cueing tasks (see Section 2.2.3) or linguistic acceptability judgment tasks (see Section 1.1.1).

Representations

The framework assumes that four different representations are necessary to process spatial relations: a perceptual representation, a conceptual representation, a reference frame, and a spatial template. The reference frame and the spatial template link the perceptual and the conceptual representation and, by extension, “map perception onto cognition and vice versa” (Logan & Sadler, 1996, p. 497).

PERCEPTUAL REPRESENTATION The perceptual representation is an “analog array of objects and surfaces” (Logan & Sadler, 1996, p. 497) and automatically created. It contains implicit perceptual information about object identities and relations between objects. Given that the information is only implicit, “further computation” (Logan & Sadler, 1996, p. 498) is necessary to extract it. The framework specifies this further computation.

CONCEPTUAL REPRESENTATION The conceptual representation consists of a spatial predicate that explicates (i) the relation (e.g., *above* or *below*), (ii) its arguments (i.e., the objects of the relation), (iii) what is the RO and what is the LO, (iv) the reference frame, and (v) the spatial template. The conceptual representation interfaces with language such that spatial prepositions can be seen as lexicalized conceptual representations. However, the (two-way) mapping between language and conceptual representation is not assumed to be always simple and straight-forward.

REFERENCE FRAME “The reference frame is a three-dimensional coordinate system that defines an origin, orientation, direction, and scale” (Logan & Sadler, 1996, p. 499). It links the perceptual representation

with the conceptual representation. The location of the origin defines which object is the RO. The orientation rotates the reference frame (e.g., according to the properties of the RO) to define where the above/below or left/right axes are. The direction distinguishes above from below or left from right. Finally, the scale sets the size of the reference frame.

SPATIAL TEMPLATE A spatial template defines spatial regions of acceptability for a relation. A spatial template is associated to a conceptual representation such that each conceptual representation has its own spatial template.

Processes

The framework assumes four processes: spatial indexing, reference frame adjustment, spatial template alignment, and computing goodness of fit.

SPATIAL INDEXING “Spatial indexing is required to bind the arguments of the relation in the conceptual representation to objects in the perceptual representation” (Logan & Sadler, 1996, p. 499). In particular, Logan and Sadler (1996) refer to the spatial index theory by Pylyshyn (1989) summarized in Section 2.1.1.

REFERENCE FRAME ADJUSTMENT This process sets the four different parameters of the reference frame (origin, orientation, direction, scale) depending on the RO. This imposes the reference frame on the RO.

SPATIAL TEMPLATE ALIGNMENT This process aligns the spatial template with the reference frame and imposes it on the RO.

COMPUTING GOODNESS OF FIT Given the acceptability regions stored in the spatial template and its alignment with the RO (which makes the regions relative to the axes of the RO), this process determines whether the location of the LO is a good, acceptable, or bad example of the spatial relation in question.

After having introduced the framework by Logan and Sadler (1996), I next present a dichotomy of visual spatial relations originating from cognitive neuroscience research. One conclusion from this research is that the size of the attentional scope affects the way how humans process spatial relations, suggesting that serial movements (or: shifts) of attention are necessary to process spatial relations.

2.2.2 Categorical and Coordinate Spatial Relations

Imagine, you are sitting in the library and writing your dissertation. To the left of your laptop, you placed your bottle of water. This spatial relation (bottle to the left of laptop) is called a ‘categorical spatial

relation'. This type of relation does not specify the exact location of the bottle, rather, it parses space into distinct categories (e.g., left or right). In the same situation, however, you might want to have a sip of water from the bottle. For this action, you (or your hand) needs precise location information to grasp the bottle. This second type of spatial relation (bottle relative to grasping hand) is called a 'coordinate spatial relation'.

In cognitive neuroscience, researchers found that these two kinds of spatial relations (categorical and coordinate spatial relations) "are processed by at least partially different underlying [neuronal] mechanisms, mainly located in the left and right hemisphere [side of the brain], respectively" (van der Ham, Postma, & Laeng, 2014, p. 142). In particular, the left hemisphere processes categorical relations better than the right hemisphere, whereas the right hemisphere processes coordinate relations better than the left hemisphere. However, this lateralization pattern can be affected by the specific task (for a recent review see van der Ham et al., 2014; see also Jager & Postma, 2003; Kosslyn, 1987, 2006)

Given that the left hemisphere is predominant in language processing and categorical spatial relations have linguistics counterparts (spatial prepositions), researchers explored whether the verbalization of categorical relations is the main factor for the left lateralization of categorical relation processing (e.g., Kemmerer & Tranel, 2000; Kranjec, Lupyan, & Chatterjee, 2014; van der Ham & Postma, 2010; see also Amorapanth et al., 2012; Kemmerer, 2006). While these studies found interactions of language and spatial relation processing², van der Ham et al. (2014, p. 145) conclude in their review: "[G]iven the current evidence it seems highly unlikely that language by itself is the determining factor in the direction of lateralization."

In a visual working memory task with simple stimuli (colored squares, Dent, 2009) as well as in a scene perception task with more complex stimuli (Rosielle, Crabb, & Cooper, 2002), changes in categorical relations were detected faster and more accurate than changes in coordinate relations. Dent (2009, p. 2372) "suggest[s] that the categorical relations are an intrinsic property of the representation of spatial configuration" in visual-spatial short-term memory (see also Olson & Marshuetz, 2005). Similarly, Rosielle et al. (2002, p. 319) "suggest that categorical spatial relations are being coded in scene perception and that attention is required in order to encode spatial relations."

Visual attention was also found to affect the performance of categorical vs. coordinate spatial relation processing. In their review, van der Ham et al. (2014) theorize that the size of the attentional scope is a main factor for the observed categorical/coordinate distinction. For instance, in the experiments conducted by Laeng, Okubo, Saneyoshi, and

The size of the attentional scope affects categorical vs. coordinate spatial relation processing.

² See Section 1.1.2 for related research in the cognitive psychology tradition that asks whether and to what extent linguistic and non-linguistic organization of space coincide.

Michimata (2011), participants first saw a spatial relation of two objects followed by a visual cue that triggered the deployment of visual attention (cf. Posner's spatial cueing paradigm, Section 2.1). Subsequently, the same objects re-appeared at the cued location and participants had to decide whether the same relation was depicted or not. The relation was either manipulated to reflect a categorical change (e.g., circle above triangle changed to triangle above circle), a coordinate change (e.g., the circle was closer to the triangle), or no change. Crucially, the visual cue either enclosed only one object (small attention window) or both objects (large attention window). Laeng et al. (2011) found that if people were cued with a small attention window, their categorical relation processing was enhanced (faster detection compared to coordinate changes). On the other hand, when cued with a large attention window, participants were quicker to detect coordinate relation changes compared to categorical relation changes (see also Franciotti et al., 2013; van der Ham et al., 2014).

Recently, Stocker and Laeng (2017) related these empirical findings with Talmy's linguistic analyses of the "windowing of attention in language" (Talmy, 2000, Chapter 4). In a similar spirit, Laeng et al. (2011, p. 322) write: "[R]esearch on the linguistic, top-down, control of attention strongly suggests the existence of diversified 'attention routines' (Ullman, 1984), which may be expressed with sequential shifts of the attention window [...]". That is, the benefit of a small attention window for processing categorical spatial relations suggests that attention serially selects each object of the relation – attention should shift from one object to the other. The idea that shifts of attention are necessary for spatial relation processing was already put forward by Kosslyn (1987, p. 170, Table 1). More recently, Franconeri et al. (2012) proposed a theoretical framework of spatial relation processing that also assumes shifts of attention.

2.2.3 *Shifting Attention to Process Spatial Relations*

Franconeri et al. (2012) discuss a variety of mechanisms that might underlie the flexibility of the visual system to process visual spatial relations. Common to all mechanisms is that the visual system needs to select the two objects that make up the spatial relation. Franconeri et al. (2012) group the mechanisms into two categories: Simultaneous selection of both objects or sequential selection, i.e., only a single object is selected at a time. Since according to Franconeri et al. (2012, p. 221), simultaneous selection "is known to bring processing difficulties associated with both object identification and binding of those identities to specific locations", they instead propose that the visual system serially selects the two objects of a spatial relation. More specifically, they suggest two mechanisms: Either, the attentional selection shifts from a global focus (encompassing both objects) to a narrow focus (selecting

To process spatial relations, humans shift their visual attention.

a single object). Or, attention first selects one of the two objects and shifts to the second object.

In support of the general hypothesis of sequential selection, Franconeri et al. (2012) present electroencephalography (EEG) evidence showing that their participants shifted spatial attention as they performed simple spatial relationship judgments. Further evidence for this ‘shift account’ of spatial relation processing is reported by Yuan, Uttal, and Franconeri (2016) and Roth and Franconeri (2012, see also Holcombe, Linares, & Vaziri-Pashkam, 2011). While Yuan et al. (2016, p. 3) try to “minimize the role of language” to investigate potential asymmetries in the perceptual representations of spatial relations (as predicted by their shift account), Roth and Franconeri (2012) explicitly investigate the role of spatial language in their framework. They claim that the linguistic asymmetry of spatial relationships (i.e., the different roles of the RO and the LO, cf. “The bike is in front of the house” vs. “The house is behind the bike”) mirrors the perceptual representation of spatial relations (see Section 1.1.3 for a summary of the experiments reported in Roth & Franconeri, 2012, as well as more research connecting linguistic and attentional processing of spatial relations).

Motivated by the same linguistic asymmetry (see Logan, 1995), Logan and Sadler (1996, p. 499, emphasis in the original) write: “The distinction between reference and located objects gives a direction to the conceptual representation; the viewer’s attention should move *from* the reference object *to* the located object (Logan 1995).” Based on Logan (1995) and Logan and Sadler (1996), Gibson and colleagues propose “a theory of how spatial symbols control the orientation of attention in space” (Gibson & Sztybel, 2014, p. 271; see Gibson & Sztybel, 2014, for a review). The studies supporting their theory are variants of Posner’s spatial cueing paradigm (see Section 2.1). For instance, Gibson and Kingstone (2006) presented either spatial words (*above/below/left/right*) or non-linguistic cues (e.g., arrows) in the center of a screen. Subjects had to report the color of a target that appeared at the cued location. This experimental paradigm presents a spatial relation (target, LO, is defined relative to a cue, RO) and it is thought that participants shift their attention from the cue to the target to perform the task (e.g., Davis & Gibson, 2012; Gibson & Kingstone, 2006; Gibson & Sztybel, 2014).

Gibson and Kingstone (2006) showed that when words were used as cues, participants were slower to respond compared to non-linguistic cues. In addition, subjects were faster in detecting targets on the vertical axis compared to the horizontal axis – but only when words were used as cues and not with non-linguistic cues. Explicitly referencing Logan (1995) and Logan and Sadler (1996), Gibson and Sztybel (2014) interpret this as an effect of the spatial reference frame (cf. Section 2.2.1) that needs to be imposed on the RO to process a linguistically described spatial relation – in contrast to attentional shifts triggered by non-linguistic cues.

Attention is theorized to move from the RO to the LO.

In a category learning task, Livins, Dumas, and Spivey (2016) showed that priming people's orientation of visual attention (with either horizontally or vertically aligned flashing circles) affected how participants categorized the stimuli (using ambiguous spatial relations – either a horizontal or a vertical oriented relation). The same type of priming also affected participants' recognition of more complex relations and analogical reasoning (i.e., mapping arguments of two depicted relations, e.g., mapping the boy to the cat in depictions of “The cat chases the mouse” and “The boy chases the cat”, Livins et al., 2016). These results support the claim that shifts of attention are an important part of relational processing. The final section of this chapter interprets the AVS model (introduced in Section 1.1.4) in terms of the summarized research on spatial relation processing as well as the conceptualizations of attention in the visual attention literature.

2.3 THE TYPE OF ATTENTION IN THE AVS MODEL

As discussed in Section 2.1.1, the term ‘attention’ refers to a variety of different concepts. In this section, I relate research on attention to the conceptualization of attention in the AVS model. Given that the authors of the AVS model unfortunately remained rather silent on this subject, this interpretation is not an “official” part of the AVS model.

In interpreting, I followed reviews about the use and role of metaphors of attention by Fernandez-Duque and Johnson (1999, 2002). Fernandez-Duque and Johnson (1999, p. 97) identify the ‘attention-as-gradient’ metaphor – a variant of the space-based ‘spotlight of attention’ metaphor – by referring to works from Downing and Pinker (1985) and LaBerge and Brown (1989, among others). These papers are also cited by Regier and Carlson (2001) when they introduce the exponential decay function of the attentional distribution implemented in the AVS model. Thus, the attentional distribution of the AVS model conceptualizes attention as a space-based spotlight with a gradual decrease from the attentional focus.

The attentional distribution in the AVS model conceptualizes attention as a gradient.

Further support for this claim comes from an AVS follow-up paper, in which Carlson et al. (2006) extend the AVS model to account for world-knowledge effects in spatial language (see Section 1.1.1). On the one hand, Carlson et al. (2006) refer to the attentional distribution in the AVS model as “an attentional spotlight” (p. 296) and as an “attentional beam” (p. 297). On the other hand, the experimental paradigm used by Carlson et al. (2006, p. 300) is perhaps even more supportive for my claim: They conducted “a speeded sentence-picture verification paradigm in which attention was cued to the left, center or right side of a rectangle by means of an exogenous cue, an established means for anchoring attention (e.g., Jonides, 1981; Posner, 1980)” – that is, they conducted a variant of Posner's spatial cueing paradigm (see

Section 2.1) which is a prominent example of conceptualizing attention as a spotlight.

The attention-as-gradient metaphor – as a variant of the spotlight metaphor – primarily deals with issues of *selecting* specific locations. In terms of the AVS model, this means that attention *selects* the RO, because its focus is defined on the RO.³ Since the RO is an object, the attentional distribution might also be seen as an instance of object-based attention. This interpretation becomes particularly interesting, if one considers the role of the attentional distribution for processing geometric object properties.

Considering the importance of attentional shifts for spatial relation processing (cf. Sections 2.2.3 and 1.1.3), the AVS model needs to do more than selecting one of two objects. More to the point, the AVS model needs to implement a shift of attention. In motivating their choice of using attention in the model, Regier and Carlson (2001) cite Logan (1994, 1995). Logan (1994, 1995) draw on the computational framework from Logan and Sadler (1996, see Section 2.2.1). Recall that this framework assumes that processing of a spatial relation starts with “spatially indexing the arguments of the relation” (Logan, 1994, p. 1015; Pylyshyn, 1989, 2001, theorize spatial indices to be pre-attentive, see Section 2.1.1). Subsequently, the framework posits that “the viewer’s attention should move from the reference object to the located object” (Logan & Sadler, 1996, p. 499). Hence, staying in this framework, the vector sum in the AVS model implements a *directed movement of attention* from the RO to the LO. This interpretation is in line with the references cited by Regier and Carlson (2001) in their motivation of the vector sum (Georgopoulos, Schwartz, & Kettner, 1986; Lee, Rohrer, & Sparks, 1988; Wilson & Kim, 1994). In particular, Georgopoulos et al. (1986) propose a vector sum representation for *movements* of (monkey) arms and Lee et al. (1988) suggest a similar representation for saccadic eye *movements* (i.e., shifts of overt attention).

The vector sum in the AVS model represents a directed movement of attention.

Interpreting the vector sum as representing a directed shift of attention also fits well into the attentional-shift account from Franconeri et al. (2012). Franconeri et al. (2012) ask about the role of shifts of attention for spatially relating *two* objects. In contrast, Regier and Carlson (2001) are primarily interested in how the geometric properties of *one* single object (the RO) affect linguistic acceptability judgments – and how these behavioral outcomes could be explained with attentional mechanisms. My interpretation of the AVS model merges these two different approaches by distinguishing the components of the AVS model: The attentional distribution selects *one* object of the spatial relation while the vector sum represents where the attentional selection should move to next (cf. Fernandez-Duque & Johnson, 1999, p. 95f.; Logan, 1995).

³ In addition, the attentional distribution is only defined on the points of the RO – strictly reading Regier and Carlson (2001, p. 277–278, Equation 10).

Taken together, I claim that the attentional distribution in the AVS model conceptualizes attention as a space-based spotlight with a gradual decrease and the directed vector sum represents a shift of attention from the RO to the LO. The direction of the attentional shift conflicts with recent evidence suggesting a shift from the LO to the RO (see Section 1.1.3). This is why I reversed the direction of the shift in the AVS model, leading to the development of the reversed AVS (rAVS) model. The following Part II of this thesis motivates and thoroughly assesses the rAVS model. In particular, the performance of the rAVS model is compared against the performance of the AVS model using existing (Chapter 3) and newly collected data (Chapter 4) as well as a variety of model comparison techniques (Chapter 5).

Part II

COMPUTATIONAL AND EMPIRICAL STUDIES

The ‘reversed AVS’ (rAVS) model is a variation of the AVS model. The main change in the computations is the reversed direction of the vector(s). Furthermore, the attentional focus in the rAVS model always lies on the LO. These changes reverse the directionality of the attentional shift: Instead of shifting from the RO to the LO (as in the AVS model), attention shifts from the LO to the RO in the rAVS model. This implements empirical findings suggesting the latter directionality of the attentional shift (see Section 1.1.3). This chapter introduces four variations of the rAVS model and evaluates them on the data from Regier and Carlson (2001).

In the rAVS models, attention moves from the LO to the RO.

In all rAVS variations, the direction of the vectors in the attentional vector sum is reversed as follows: Instead of being rooted at every point in the RO and pointing to the LO, the vectors are rooted at every point in the LO and point to one particular point in the RO. This particular point in the RO must be defined. By defining different points, different variations of the rAVS model emerge. In this chapter, I present and evaluate four variations of the rAVS model that differ in their vector end point. The direction of the reversed vector sum is finally compared to canonical downwards instead of canonical upright (in the case of *above*). “This flip [of reference direction] is counterintuitive, but certainly not computationally difficult” (Roth & Franconeri, 2012, p. 7). The height component of the AVS model is not changed in the rAVS model. As in the AVS model, it takes the y-value of the LO as input and computes the height according to the top of the RO (see Equation 1.1 on page 13 and Figure 1.1a). The following formulas describe the rAVS model mathematically:

$$\overrightarrow{\text{vectorSum}} = \sum_{i \in \text{LO}} a_i \cdot \overrightarrow{i, \text{R}} \quad (3.1)$$

$$\delta = \angle(\text{down}, \overrightarrow{\text{vectorSum}}) \quad (3.2)$$

$$\text{above}(\text{LO}, \text{RO}) = g(\delta) \cdot \text{height}(\text{LO}_y) \quad (3.3)$$

The variables and functions here are the same as for the AVS model (see Equations 1.1–1.7 on pages 13–15) – except for changes regarding the reversal of the attentional shift: i denotes a single point of the LO, a_i denotes the amount of attention at LO’s point i , and $\overrightarrow{i, \text{R}}$ describes the

* Parts of the work presented in Chapter 3 were published in Kluth, Burigo, and Knoeferle (2015, 2016c, 2017). However, the published papers neither report any other rAVS variation than the rAVS_{w-comb} model nor do they present a detailed model evaluation on the level of individual experiments.

vectors pointing from all points i of the LO to one particular point R of the RO – to which point exactly depends on the specific rAVS variation. In contrast to the AVS model, the direction of $\vec{\text{vectorSum}}$ (the final vector) is compared to canonical downwards (denoted down) instead of canonical upwards to obtain the deviation as argument for the $g()$ function.

The LO is simplified as a single point.

ATTENTIONAL DISTRIBUTION In much spatial language literature, the LO is simplified to consist of a single point only. To be able to compare the rAVS variations with the AVS model I stick to this convention in this project. Thus, the LO remains a single point in the rAVS models and the location of the focus is always well-defined: it is at the same location as the LO. Due to the simplification of the LO as a single point, the following implications regarding the role of the attentional distribution emerge for all rAVS variations. The amount of attention at the single-point LO always equals 1 ($a_0 = 1$) because the single point of the LO coincides with the attentional focus. Moreover, the attentionally weighted vector sum consists of only one single vector. Even a different amount of attention at the vector root (i.e., $a_0 \neq 1$) would not affect the final rating (as long as $a_0 > 0$). This is because neither the AVS nor the rAVS models consider the length of the vector sum in their angle computation. Thus, the attentional distribution (and hence the parameter λ controlling it) does not have an impact on the outcome of the rAVS models. This limitation is only valid for simplified LOs. Future research should investigate the role of extended LOs for the rAVS models.

3.1 MOTIVATING RAVS VARIATIONS

Proximal orientation and center-of-mass orientation affect human acceptability ratings.

Previous research suggests that the angle between the RO and the LO is an important factor in judging the acceptability of a spatial preposition (Gapp, 1995; Hayward & Tarr, 1995; Regier, 1996; Regier & Carlson, 2001). More specifically, in empirically assessing the AVS model, Regier and Carlson (2001) found that the orientation of two imaginary lines affected human acceptability ratings: the ‘proximal orientation’ and the ‘center-of-mass orientation’ (see also Regier, 1996). The proximal orientation is the orientation of the line that connects the LO with the proximal point P on the RO (see loosely dashed line in Figure 3.1b). The center-of-mass orientation is the orientation of the line that connects the LO with the center-of-mass of the RO (see solid line in Figure 3.1b).

The AVS model is able to compute either of these orientations or combinations of both using different magnitudes of its attentional width (see Regier & Carlson, 2001, p. 278 and appendix). If the attentional width is maximal (i.e., the attentional distribution is of uniform strength), the whole RO receives the same amount of attention and the AVS model computes the center-of-mass orientation. If the attentional

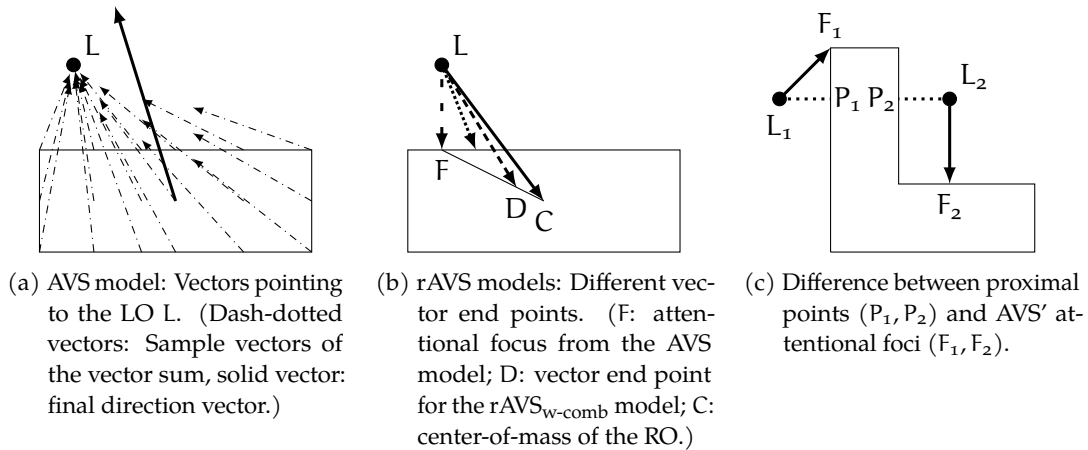


Figure 3.1: Vector end points in (a) the AVS model and (b) the rAVS models ($rAVS_{prox}$: loosely dashed, $rAVS_{comb}$: dotted, $rAVS_{w-comb}$: densely dashed, $rAVS_{c-o-m}$: solid). The points F_1 and F_2 in (c) are AVS' attentional foci for LOs L_1 and L_2 . These are used in the $rAVS_{prox}$ model (solid vectors) as well as in the $rAVS_{comb}$ and $rAVS_{w-comb}$ models.

width is minimal¹, only one point of the RO (the attentional focus F) is attended. Given that F and the proximal point P often coincide, this means that the AVS model computes the proximal orientation with its minimal attentional width. Intermediate values of the attentional width lead to different weighted averaging of the two orientations.

Accounting for the two orientations while reversing the direction of the attentional shift has led to four different rAVS variations. Two of these variations use only one orientation: the $rAVS_{c-o-m}$ model uses the center-of-mass orientation only and the $rAVS_{prox}$ model uses the proximal orientation only. The other two rAVS variations combine the two orientations in different ways: the $rAVS_{comb}$ model computes the mean of the two orientations and the $rAVS_{w-comb}$ model weights the influence of the two orientations using 'relative distance'. Relative distance is defined to be absolute distance divided by the dimensions of the RO (more details below). Its definition was informed by the empirical data from Regier and Carlson (2001, exp. 7).

All rAVS variations differ only with respect to how they compute the vector end point R in Equation 3.1. In the AVS model, the vector end point is unambiguous: It is the LO, simplified to be a single point (shown as circle labeled with L in Figure 3.1a). The vector sum in the rAVS variations, however, points to the RO which consists of more than one point. This is why the vector end point R must be determined.

¹ but greater than zero

rAVS_{C-O-M} MODEL In the rAVS_{C-O-M} model, the vector end point R is defined to be the center-of-mass C of the RO. The orientation of the vector sum equals the center-of-mass orientation (see solid arrow in Figure 3.1b). Apart from research on spatial language highlighting the importance of the center-of-mass orientation, the use of the center-of-mass is also in line with findings from saccadic localization. If humans are instructed to “look at a object as a whole” their first fixation (in most cases) lands on the center-of-mass (Brouwer, Franz, & Gegenfurtner, 2009; Melcher & Kowler, 1999; Vishwanath & Kowler, 2003; see also Section 2.1.1).

rAVS_{PROX} MODEL In the rAVS_{prox} model, the vector sum points to the proximal point F on the RO (see loosely dashed vector in Figure 3.1b). This equals the proximal orientation, if the LO consists of a single point. But where is the proximal point? The basic definition is: The point on the RO that has the lowest distance to the LO. Applying this definition, we obtain the following proximal points: If the LO is above the ‘grazing line’² of the RO, the proximal point is always on top of the RO. If the LO, however, is, say, to the right of the RO, the proximal point is at the right side of the RO. Considering an RO as depicted in Figure 3.1c, the proximal orientations of the two LOs L₁ and L₂ (the dotted lines) result in the same deviation. That is, for this example the rAVS_{prox} model would compute an equal rating for both LOs L₁ and L₂. Not surprisingly, simulations of the rAVS_{prox} model with this definition of proximal points do not result in good fits to empirical data. This is why I am not using the proximal point in the literal sense for the vector end point of the rAVS_{prox} model. Rather, the vector end point in the rAVS_{prox} model is always the same as the location of the attentional focus F in the AVS model (in Figure 3.1c, points F₁ and F₂ for LOs L₁ and L₂, respectively). That is, the vector end point is determined by letting fall a perpendicular from the LO to the RO or – if the LO is to the left or right of the RO – choosing the closest point on top of the RO. For Figure 3.1c, this results in the solid vectors. Accordingly, LO L₁ now gets a lower rating than LO L₂.

rAVS_{COMB} MODEL Since there is evidence that both the center-of-mass orientation and the proximal orientation are important for the comprehension of spatial language, both orientations are averaged in the rAVS_{comb} model (see dotted vector in Figure 3.1b). If both orientations are averaged, both orientations must be computed first. For instance, in Figure 3.1b, the loosely dashed vector (proximal orientation) and the solid vector (center-of-mass orientation) must be computed to be able to obtain the dotted vector. Thus, to account for both orientations two vector sums must be calculated. Since every vector sum only con-

² The grazing line is an imaginary horizontal line that touches the top-most point of the RO, see Regier & Carlson, 2001, exps. 5 & 6, reviewed in detail in Section 3.2.5.

sists of one vector (due to the single point LO) this should not lead to much higher cognitive workload and is thus a plausible possibility. The averaging in the $rAVS_{comb}$ model is done with the following formula:

$$\text{angle}_{combined} = \frac{\text{angle}_{prox} + \text{angle}_{c-o-m}}{2} \quad (3.4)$$

Note that the angle_{prox} is based on the modified definition of a proximal point described above for the $rAVS_{prox}$ model (i.e., the $rAVS_{comb}$ model follows the $rAVS_{prox}$ model and uses the location of the attentional focus F in the AVS model instead of the literal proximal point).

$rAVS_{w-comb}$ MODEL The $rAVS_{comb}$ model always takes the average deviation of the two deviations from the $rAVS_{c-o-m}$ and the $rAVS_{prox}$ models. That is, both orientations are of equal importance and this importance is the same for all LOs. As we will see later, there is evidence that the relative importance of the two orientations is not fixed for all possible locations of LOs (experiment 7 from Regier & Carlson, 2001, see Section 3.2.6). This is why I have developed a fourth variation of the rAVS model: the $rAVS_{w-comb}$ model.

In this model, the vector end point R lies on the imaginary line that connects the center-of-mass C of the RO with the proximal point F (see point D in Figure 3.1b). Again, the modified definition of proximity is used, i.e., F is the same as the attentional focus in the AVS model (see Figure 3.1c). The location of the vector end point D depends on the ‘relative distance’ of the LO to the RO: For distant LOs, D is closer to C; for close LOs, D is closer to F. This means that for distant LOs the center-of-mass orientation is more important than for close LOs. In contrast, the proximal orientation is more important for close LOs than for distant LOs. The distance of an LO is hereby considered in relative terms. That is, even if the absolute distance between an RO and an LO remains equal, the dimensions of the RO (i.e., width and height) affect the relative distance. The relative distance between an LO and an RO is computed as follows:

$$\text{dist}_{rel.}(LO, RO) = \frac{|LO, P|_x}{RO_{width}} + \frac{|LO, P|_y}{RO_{height}} \quad (3.5)$$

Here, $|LO, P|_x$ denotes the horizontal component of the absolute distance between the LO and the proximal point P on the RO while the corresponding vertical component is denoted as $|LO, P|_y$.

Note that P is the literal proximal point which is different from the attentional focus F. P is the point on the RO that has the smallest absolute distance to the LO, regardless of the shape of the RO. In contrast, the attentional focus F is defined to lie on one specific side of the RO only (e.g., on the top of the RO if the preposition is *above*). P and F coincide if the proximal point happens to be on the same side as the attentional focus (e.g., if P is on the top of the RO and the preposition

is *above*; compare also the different locations of points P_1, P_2 with F_1, F_2 in Figure 3.1c).

The relative distance is combined with the additional free model parameter α to compute the location of the vector end point D. More specifically, this is realized with the following linear function:

$$D = \begin{cases} \overrightarrow{LO, C} + (-\alpha \cdot \text{dist}_{\text{rel.}} + 1) \cdot \overrightarrow{C, F} & \text{if } (-\alpha \cdot \text{dist}_{\text{rel.}} + 1) > 0 \\ C & \text{else} \end{cases} \quad (3.6)$$

3.1.1 Comparison to PC(-BB) Models from Regier and Carlson (2001)

As competitor models to the AVS model, Regier and Carlson (2001) suggested three other models: The BB model (bounding box model), the PC model (proximal and center-of-mass model), and the PC-BB model (a combination of the PC and the BB model). How are the various rAVS models related to these other models? Before answering this question, I briefly introduce the three models:

BB MODEL Regier and Carlson (2001) define the BB model in the following way:

“According to this [BB] model, a trajector object [LO] is above a landmark object [RO] if it is higher than the highest point of the landmark and between its rightmost and leftmost points.” (Regier & Carlson, 2001, p. 274)

To achieve this rating, the BB model consists of the same height component as the AVS model for the vertical component (see Equation 1.1 on page 13) and a combination of two sigmoid functions for the horizontal component:

$$\text{center}(x) = \text{sig}(x - \text{left}, \text{lrgain})^{\text{lrex}} \cdot \text{sig}(\text{right} - x, \text{lrgain})^{\text{lrex}} \quad (3.7)$$

The sig() function is defined in Equation 1.2 (see page 1.2), lrgain and lrex are free parameters of the BB model. Both components (horizontal and vertical) are multiplied to obtain the final rating:

$$\text{above}(LO_x, LO_y) = \text{height}(LO_y) \cdot \text{center}(LO_x) \quad (3.8)$$

PC MODEL In contrast to the BB model, the PC model uses angular deviations, namely the center-of-mass and proximal orientations:

“Formally, the PC model characterizes above as a linear combination of the degrees of alignment of the center-of-mass

and proximal orientations with upright vertical (Regier, 1996, 1997) as in Equation [3.9]:

$$\text{above} = \alpha f(\text{com}) + (1 - \alpha) f(\text{prox}) \quad (3.9)$$

(Regier & Carlson, 2001, p. 276)

Here, *com* and *prox* describe the proximal and center-of-mass orientations, α is a free parameter of the PC model that weights the importance of each orientations and $f()$ is a function that maps angular deviation to rating:

$$f(\text{angle}) = (\text{slope} \cdot \text{angle} + \text{y-intercept}) \cdot \text{sig}(90 - \text{angle}, \text{gain}) \quad (3.10)$$

The PC-model does not contain an explicit height component. The sigmoid part in the $f()$ function can be interpreted as a functionally similar part, as it results in low ratings for angular deviations greater than 90 degrees (see also Figure 3 in Regier & Carlson, 2001). However, the grazing line is not explicitly formulated in the PC model (see Regier & Carlson, 2001, exps. 5 & 6, reviewed in detail in Section 3.2.5).

PC-BB MODEL The PC-BB model is a combination of the PC and the BB model. Basically, it includes the missing height component in the PC model. The height component from the BB model (which is also the same in the AVS model, see Equation 1.1 on page 13), is multiplied with the relative importance of center-of-mass and proximal orientation (from the PC model):

$$\text{above}(\text{LO}) = \text{height}(\text{LO}_y) \cdot [\alpha \cdot g(\text{com}) + (1 - \alpha) \cdot g(\text{prox})] \quad (3.11)$$

Another difference to the PC model is the use of another function that maps angular deviation to ratings. Instead of the function $f()$ (see Equation 3.10), the simpler function $g()$ is used:

$$g(\text{angle}) = \text{slope} \cdot \text{angle} + \text{y-intercept} \quad (3.12)$$

The same function is also used in the AVS model (see Equation 1.6). The sigmoid part in function $f()$ used by the PC model is now incorporated in the height component that comes from the BB model.

COMPARISON The BB model underlies a different assumption compared to the rAVS models. The rAVS models are using angular deviations, i.e., polar coordinates, whereas the BB model operates with Cartesian coordinates (see also Regier & Carlson, 2001, p. 275). However, both the PC and PC-BB models use angular features, too. The main difference between these two models is that the PC-BB model explicitly accounts for the grazing line with its height component, whereas the

PC model does not. In fact, Regier and Carlson (2001) showed that the PC model cannot accommodate effects of the grazing line, if the two orientational features are held constant (experiments 5 and 6, Regier & Carlson, 2001, see Section 3.2.5). All rAVS models contain the same height component as the BB, the PC-BB, and the AVS model – thus, all rAVS models should be able to accommodate effects of the grazing line.

The $rAVS_{\text{prox}}$ and the $rAVS_{\text{c-o-m}}$ model differ from the PC(-BB) models, because they consider only one orientation instead of both at the same time. The $rAVS_{\text{comb}}$ model averages both orientations, but does this with a fixed averaging formula, whereas the PC(-BB) models have a free parameter α that can be adjusted to the data. However, this parameter is valid for all LOs at the same time. If, say, $\alpha = 0.3$, then the center-of-mass orientation only contributes with 30% to the final rating, but the proximal orientation contributes with 70% – for *all* LOs.

The $rAVS_{\text{w-comb}}$ model, however, is able to apply different proportions of importance within one set of parameter, depending on the relative distance of the LO. This is the main crucial difference between the $rAVS_{\text{w-comb}}$ model and the PC(-BB) model. Thus, in the $rAVS_{\text{w-comb}}$ model, the proximal orientation is more important for close LOs than for distant LOs, whereas the center-of-mass orientation is more important for distant LOs than for close LOs – with a fixed set of parameters.

The $rAVS_{\text{w-comb}}$ also uses two different interpretations of a proximal point. For the computation of the relative distance, the closest point on the RO is used. For the computation of the vector end point, however, the modified definition of proximal point is used: This point always lies on top of the RO and is the same as the attentional focus in the AVS model.

The AVS model can be interpreted as combining the two orientations via the vector sum. Since the vector sum is weighted by attention, which in turn is influenced by the distance of the LO, the AVS model also applies different importances of the two orientations to different LOs within one set of parameters. The vector sum makes the AVS model flexible (accounting for the geometry of objects and weighting center-of-mass and proximal orientation) but also computationally expensive. Due to the simplification of the LO, the rAVS model variations are using only a single vector.³ Does any of the rAVS variations accommodate the same empirical effects as the AVS model via its vector sum? To answer this question, I have evaluated the four rAVS variations with the same empirical data that Regier and Carlson (2001) used to evaluate the AVS model (the data from Regier & Carlson, 2001).

3.2 MODEL EVALUATION

The AVS model and the four rAVS model variations have either four (AVS, $rAVS_{\text{prox}}$, $rAVS_{\text{c-o-m}}$, $rAVS_{\text{comb}}$) or five ($rAVS_{\text{w-comb}}$) free param-

³ or only two vectors in the $rAVS_{\text{comb}}$ model

ters. These free parameters allow the models to have a certain flexibility in their output. To facilitate the comparison of the two models, the rAVS variations were designed to be as close as possible to the AVS model. This is why all four parameters of the rAVS_{prox}, rAVS_{c-o-m}, and rAVS_{comb} models and four out of five parameters of the rAVS_{w-comb} model are the same as in the AVS model. These four parameters are the highgain parameter (used in the model component that adapts the score based on the vertical location of the LO), the slope and intercept parameters (slope and intercept of the linear function that maps angular deviations to acceptability scores), and the parameter λ that controls the width of the attentional distribution. However, note that the attentional distribution (and hence the parameter λ) does not affect the outcome of the rAVS models with single-point LOs (see page 34). Additionally, the rAVS_{w-comb} model specifies the strength of the relative distance on the vector sum direction with its parameter α . In the following, I assess all models on the data from Logan and Sadler (1996), Hayward and Tarr (1995), and Regier and Carlson (2001) by identifying appropriate values of these free model parameters. The first model benchmark is the ‘goodness-of-fit’ (GOF) value and the second model benchmark is the ‘simple hold-out’ (SHO) value. The next section introduces these two measures.

Contrasting implementations of the attentional shift are assessed by measuring model performance on empirical data.

3.2.1 Goodness-of-Fit and Simple Hold-Out: Method

Goodness-of-Fit

To compute the GOF value, the ‘normalized Root Mean Square Error’ (nRMSE) is minimized, i.e., the values of all free model parameters are estimated to get the tightest fit to the empirical data. The nRMSE is defined as follows, with N being the number of data points:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_i^N (\text{data}_i - \text{modelOut}_i)^2} \quad (3.13)$$

$$\text{nRMSE} = \frac{\text{RMSE}}{\text{rating}_{\max} - \text{rating}_{\min}} \quad (3.14)$$

The GOF provides information on how well a model can simulate empirical data. If a model is not able to fit the data, there is no need to further consider this model. The closest possible fit has an nRMSE of zero – the output of the model equals the empirical data at every data point. The worst possible fit has an nRMSE of 1.0 (the RMSE normalized with the range of possible ratings).

To estimate the parameters of a model, I applied the ‘simulated annealing’ method. Simulated annealing is a special case of the Metropolis algorithm, named after Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953). The Metropolis algorithm is a ‘Markov Chain Monte Carlo’ approach (MCMC, for more theoretical background see Madras,

2002). The simulated annealing algorithm samples new parameter values in the vicinity of the current parameter values (using a Gaussian distribution as proposal distribution) and evaluates the fit of the model with these new parameters. Depending on whether the new fit is better or worse (compared to the fit with the previous parameters), the algorithm accepts the new parameters and uses them in the next iteration as the mean of the Gaussian distribution to sample new parameters. The algorithm reduces the standard deviation of the Gaussian proposal distribution every fixed number of iterations, such that in the beginning, the parameter search is broad – newly sampled parameter values have a high chance to be relatively distant from the current values – and in the end, the algorithm fine-tunes the parameter values – newly sampled parameter values are close to the current ones. In comparison to simple gradient descent methods, simulated annealing has the advantage to not get stuck in local minima by temporarily accepting worse parameter values. The procedure of simulated annealing in pseudo-code is given in Algorithm 1. I tested a range of different values for the parameters of the algorithm and found that the following values performed best: temperature = 0.25, iterations = 500, cooling_period = 300. I initialized the parameters accordingly.

If not stated otherwise the starting values of the model parameters were set to the parameters reported by Regier and Carlson (2001, their Table 1, Logan & Sadler, 1996, fit). The model parameters were constrained as follows:

$$-1/45 \leq \text{slope} \leq 0 \quad (3.15)$$

$$0.7 \leq \text{intercept} \leq 1.3 \quad (3.16)$$

$$0 \leq \text{highgain} \leq 10 \quad (3.17)$$

$$0 < \lambda \leq 5 \quad (3.18)$$

$$0 < \alpha \leq 5 \quad (3.19)$$

Comparing the quality of models solely on their ability to closely fit data is problematic (Pitt & Myung, 2002; Roberts & Pashler, 2000). A good fit to empirical data is necessary for a “good” model of cognitive processes. However, it is not sufficient for a thorough model evaluation. This is why I have applied the simple hold-out method (Schultheis et al., 2013) as a complementary method to assess the models.

Simple Hold-Out

One specific problem of the GOF is that it is agnostic to the source of variation in the empirical data (Pitt & Myung, 2002). I am interested in the systematic variation in the data that I can attribute to the different conditions of the task but not to the random variation of the data (i.e., noise in the data). If one model fits data better than another model, one cannot know if this is due to a better approximation of the systematic

Algorithm 1: Simulated annealing algorithm in pseudo-code.

```

for each parameter  $p$  do
  |  $p$  = starting value;
end
old_nRMSE = compute_nRMSE(data, parameters);
best_nRMSE = old_nRMSE;
for  $i = 1$  to iterations do
  | for  $j = 1$  to cooling_period do
    | for each parameter  $p$  do
      | // sample new parameter value close to actual
      | value
      | new_p = sample_Gaussian_distribution( $\mu = p$ ,
      |  $\sigma = \text{temperature}$ );
    | end
    | new_nRMSE = compute_nRMSE(data,
    | new_parameters);
    | if new_nRMSE < old_nRMSE then
      | // better result, accept parameter values
      | for each parameter  $p$  do
      | |  $p = \text{new\_p}$ ;
      | end
      | if new_nRMSE < best_nRMSE then
      | | best_nRMSE = new_nRMSE;
      | end
    | else
      | // worse result but still accept with some
      | probability
      | accept = sample_uniform_distribution(0, 1);
      | if accept  $\leq$ 
      | exp( $-(\text{new\_nRMSE} - \text{old\_nRMSE})/\text{temperature}$ )
      | then
      | | for each parameter  $p$  do
      | | |  $p = \text{new\_p}$ ;
      | | end
      | end
    | end
    | old_nRMSE = new_nRMSE;
  | end
  | // cool down
  | temperature = 0.99 · temperature;
end

```

variation or due to a closer fit to the noise in the data (the latter is known as ‘over-fitting’ data). Some models might fit noisy data better because they are more flexible than other models. Model flexibility here means the ability of a model to generate different data sets. The more different

data sets a model can generate, the more flexible is the model. One way to control for this flexibility is the evaluation of the generalizability of the models' output (Pitt & Myung, 2002). If the output of a model that was fitted to one data set does not generalize well to a different data set it was not fitted to, the model most probably over-fitted the data (i.e., obtained a better fit without explaining the systematic variation).

The simple hold-out (SHO) method controls for the generalizability of models by evaluating their ability to generalize to unseen data points. The SHO method performs well compared to other methods of model comparison (Schultheis et al., 2013). The SHO is a cross-validation method. Cross-validation is a widely used method (with several variants) to avoid over-fitting for computational models (e.g., Arlot & Celisse, 2010). The key idea of cross-validation (and thus also of SHO) is to use only a part of the data to estimate parameters (or: train the model) and to use these parameters to "predict" the remaining data (or: test the model). This is done several times using different partitions of the data. For each iteration, the nRMSE of the "prediction" (the GOF to unseen data) is saved. In the end, the median of all prediction errors is used as an evaluation measure. The lower this median prediction error, the better the model is able to generalize to unseen data. Algorithm 2 shows this procedure in pseudo-code. I have used the following parameter values: `amount_of_training_data = 70%`, `iterations = 101`.

The SHO method as proposed in Schultheis et al. (2013) does not account for cases in which the medians of the prediction errors of two competing models are almost similar – it always considers the model with the lowest median prediction error as the better model. However, the computation of the prediction errors contains random sampling (splitting the data and estimating parameters). Thus, the prediction errors are also subject to random variation. The magnitude of this randomness can be measured with the confidence interval of the median prediction error. Accordingly, I also report 95% confidence intervals of the median prediction errors (cf. Cumming, 2014). Specifically, I have used the R package `boot` (Canty & Ripley, 2016; Davison & Hinkley, 1997) with 100,000 bootstrap samples to estimate the BCa confidence intervals.

Although the SHO method gives good results without such confidence intervals (as shown by Schultheis et al., 2013), these intervals will prove useful if two models are virtually identical in their performance. As a double check to see whether the *median* of the prediction error (as used in Schultheis et al., 2013) is an appropriate evaluation measure, I also computed the *mean* of the prediction errors with corresponding BCa bootstrap confidence interval. In order to compare the AVS model with the different rAVS model variations, I computed GOF and SHO values for the data from Logan and Sadler (1996) and Hayward and Tarr (1995, Section 3.2.2) as well as the data from Regier and Carlson (2001, Sections 3.2.3–3.2.7).

Algorithm 2: Simple hold-out algorithm in pseudo-code.

```

for  $i = 0$  to iterations do
  training_data =
    pick_random(all_data, amount_of_training_data);
  fitted_parameters =
    parameter_estimation(training_data);
  test_data = all_data – training_data;
  prediction_errors[i] =
    compute_nRMSE(test_data, fitted_parameters);
end
return median(prediction_errors)

```

3.2.2 *Logan and Sadler (1996, Exp. 2, Above) and Hayward and Tarr (1995, Exp. 2, Above)*

I have fitted the acceptability rating data from Logan and Sadler (1996, exp. 2, *above*) and Hayward and Tarr (1995, exp. 2, *above*) with the AVS model and the different variations of the rAVS model.⁴ In these experiments, participants had to judge the acceptability of the spatial preposition *above* given a two-dimensional spatial configuration of one RO and one LO (similar to the experiments from Regier & Carlson, 2001). Figure 3.2 shows the positions of the RO and the LOs for these experiments. In the background of Figures 3.2b and 3.2d, spatial templates (cf. Section 2.2.1) are depicted (in which lighter colors code higher ratings). These spatial templates were computed with the rAVS_{w-comb} model using its best-fitting parameter values for the corresponding data set.

Notation: [Authors (year, exp. X, preposition)] codes for acceptability ratings for [preposition] presented as [experiment X] in [Authors (year)].

Results for Logan and Sadler (1996, Exp. 2, Above)

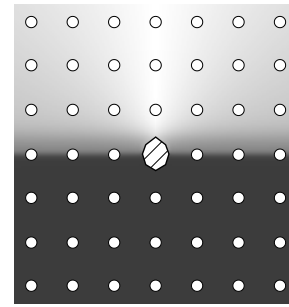
Figure 3.3a shows the results for fitting the data from Logan and Sadler (1996, exp. 2, *above*). As evident from the small GOF values, all models closely fit the data (remember that the worst nRMSE is 1.0). So, every model passed the GOF test and none of the models fits the data considerably better than any of the other models. But how do the models generalize to unseen data? This question is addressed by the SHO results which are also shown in Figure 3.3a. The SHO values are all close to the GOF values, indicating a neglectable influence of over-fitting. More importantly, all SHO results are very similar to each other and all 95% confidence intervals overlap considerably. Thus, these results do not favor any of the models.

Table 3.1 shows the parameters that gave the closest fit to the data from Logan and Sadler (1996, exp. 2, *above*). The nRMSE as well as correlation coefficients are displayed. The high correlation coefficients (for all models higher than 0.975) support the conclusion that all models closely fit the data.

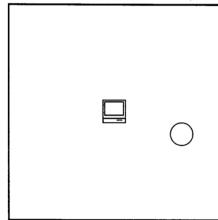
⁴ The documented source code for all model simulations reported in this thesis is available from Kluth (2018).

X	X	X	X	X	X	X
X	X	X	X	X	X	X
X	X	X	X	X	X	X
X	X	X	O	X	X	X
X	X	X	X	X	X	X
X	X	X	X	X	X	X
X	X	X	X	X	X	X

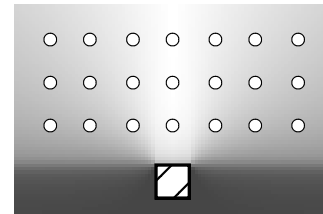
(a) Reconstruction of layout of experimental display of Logan and Sadler (1996, exp. 2). O: RO, Xs: LOs. Only one LO was shown in one trial. The grid was invisible. Size of cells in original display: width 1.25 cm, height 1.33 cm.



(b) Display used for simulating the stimuli from Logan and Sadler (1996, exp. 2, *above*).



(c) Sample experimental display of Hayward and Tarr (1995, exp. 2). The circle (LO) was shown on 48 positions around the computer (RO) in an invisible grid. A second set of RO and LO was used (floating raft and bird/fish). Image copyright: See Appendix E.



(d) Display used for simulating the stimuli from Hayward and Tarr (1995, exp. 2, *above*).

Figure 3.2: Layout of experimental displays and displays used for model simulations for (a, b) Logan and Sadler (1996, exp. 2, *above*) and for (c, d) Hayward and Tarr (1995, exp. 2, *above*). For (b, d): LOs are displayed as circles for visualization purposes. The simulations used single point LOs (at the center of the circles). Backgrounds depict $rAVS_{w-comb}$'s spatial template (lighter color coding higher rating) computed with best fitting parameters for the corresponding data set. For (d): No measurements were reported in Hayward and Tarr (1995), so the same distances as for the Logan and Sadler (1996) data were used. Only LO positions above the RO are considered, because *above* ratings for positions below the LO were not reported.

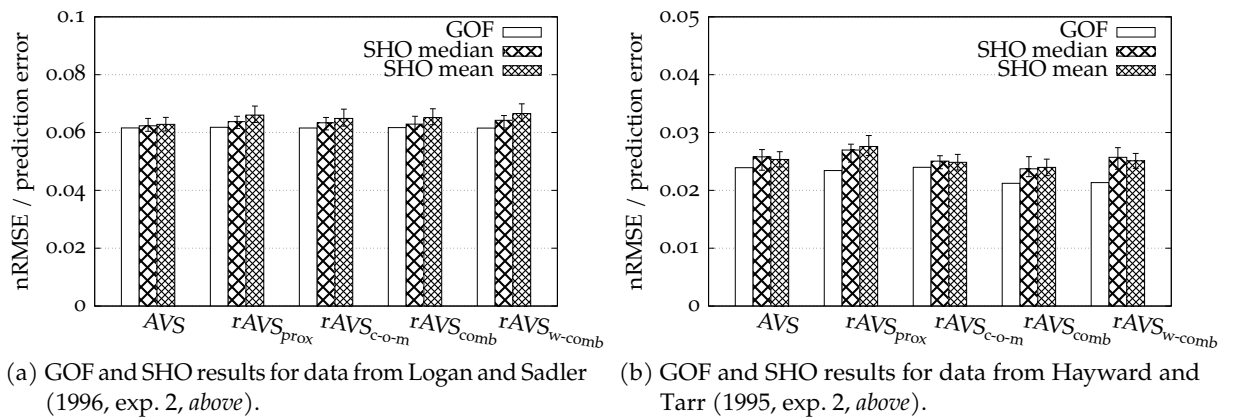


Figure 3.3: GOF and SHO results for data from (a) Logan and Sadler (1996, exp. 2, above) and (b) Hayward and Tarr (1995, exp. 2, above). Note the different y-axes. Error bars depict 95% confidence intervals of SHO median or mean respectively.

Results for Hayward and Tarr (1995, Exp. 2, Above)

The results for the data from Hayward and Tarr (1995, exp. 2, above) should be considered carefully because Hayward and Tarr (1995) did not report the exact measurements of their experimental displays. Given that their experiment is closely related to experiment 2 from Logan and Sadler (1996), I have used the same measurements (see Figure 3.2d). That said, Figure 3.3b shows the results for fitting the data from Hayward and Tarr (1995, exp. 2, above). First of all, all models fit the data even closer than the data from Logan and Sadler (1996, see lower nRMSE-range compared to Figure 3.3a). This might be because the *above* rating data from Hayward and Tarr (1995) consist only of LOs placed above the RO and no LOs placed below the RO (see Figure 3.2d). Participants judged all these LOs to be good examples of *above* (i.e., they gave high ratings). This has possibly reduced the variety of ratings in the data set and accordingly enhances model fitting performance.

The AVS model gets the worst GOF among all models; the rAVS models fit equally well. However, due to the overall small nRMSE, this difference is not important. Considering the SHO values, the rAVS_{comb} model obtains the best result. All confidence intervals, however, overlap with each other which renders this advantage inconclusive.

Interim Discussion

So far, the presented results show a comparable performance of the AVS model and all rAVS variations. Accordingly, I conclude that the direction of the attentional shift as implemented in the AVS model is not essential to accommodate the results from Logan and Sadler (1996) and Hayward and Tarr (1995). Whether attention shifts from the RO to

All rAVS variations accommodate results from Logan and Sadler (1996) and Hayward and Tarr (1995).

Table 3.1: Parameter values, correlation coefficients, and nRMSE of best fits to data from Logan and Sadler (1996, exp. 2, *above*). λ values for rAVS models are presented in parentheses because they do not change the model outcome (see page 34).

Model	λ	slope	intercept	highgain	R ²	adj. R ²	nRMSE
AVS (fit from Regier & Carlson, 2001)	1.000	-0.006	1.007	0.131	0.977	0.977	0.066
AVS (my best fit)	0.264	-0.004	0.945	0.243	0.985	0.985	0.062
rAVS _{prox}	(1.221)	-0.005	0.935	0.213	0.985	0.984	0.062
rAVS _{c-o-m}	(0.540)	-0.005	0.955	0.338	0.986	0.986	0.062
rAVS _{comb}	(2.987)	-0.005	0.943	8.882	0.987	0.987	0.062
rAVS _{w-comb}	(2.000)	-0.005	0.952	0.274	0.986	0.985	0.062
rAVS _{w-comb}	$\alpha = 1.572$						

the LO (AVS model) or whether it shifts the from LO to the RO (rAVS variations) – the ability of the models to fit the data is not impacted.

However, the different variations of the rAVS model also performed equally well. Thus, with these results none of the four rAVS variations can be favored over the others. That is, it cannot be answered where the end point of the vector in the rAVS model should be. One reason for this is that Hayward and Tarr (1995) and Logan and Sadler (1996) used comparably small ROs. With small ROs the difference between the proximal orientation and the center-of-mass orientation is small and thus it does not strongly affect the output of the model. Using larger ROs, Regier and Carlson (2001) explicitly tested the influence of these two orientations. Given that the different rAVS variations are based on these two orientations, the data from Regier and Carlson (2001) should lead to a distinct performance of the model variations.

The larger ROs by Regier and Carlson (2001) provide a promising testbed.

In the following sections, the GOF, SHO, and qualitative fit results for the data from the seven experiments reported in Regier and Carlson (2001) are presented.⁵ Regier and Carlson (2001) discussed their results in four subsets, based on the tested effects. My simulations follow this division. Apart from reporting the performances of the models on the empirical data as measured by the GOF and SHO, I am following Regier and Carlson (2001) and am also reporting the correlation of empirical data and model-generated data. For generating these data with the models, I have used the parameter values of the best fit to the data from Logan and Sadler (1996, exp. 2, *above*, see Table 3.1) – again following Regier and Carlson (2001). Thus, the correlations provide information about how well the models account for the data from Regier and Carlson (2001) without being fitted to these data.

⁵ I thank Terry Regier and Laura Carlson for sharing their data.

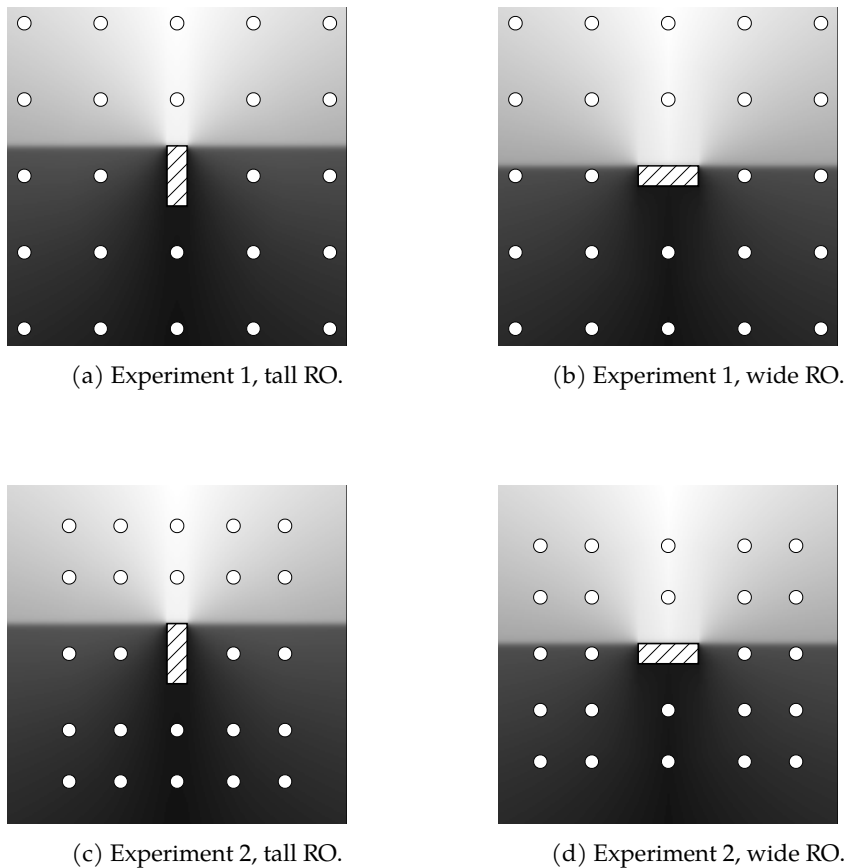


Figure 3.4: Displays used for simulating the stimuli from (a, b) exp. 1 and (c, d) exp. 2 from Regier and Carlson (2001, reconstructed from data provided by Regier & Carlson, 2001). LOs are displayed as circles for visualization purposes. The simulations used single point LOs (at the center of the circles). For critical manipulation see Figure 3.5b. ROs are patterned for visualization purposes only. Backgrounds depict $rAVS_{w-comb}$'s spatial templates (lighter color coding higher rating) computed with best fitting parameters for data from Regier and Carlson (2001, exps. 1–3).

3.2.3 *The Effect of Proximal and Center-of-Mass Orientation: Regier and Carlson (2001, Exps. 1–3)*

The first three experiments from Regier and Carlson (2001) were designed to investigate the influence of proximal orientation and center-of-mass-orientation. To this end, one of these orientations was held constant while the other orientation was varied. In the first two experiments, the same two ROs (tall and wide rectangle) were used but the placements of the LOs were manipulated. This implies a manipulation of center-of-mass or proximal orientation across the two experiments. Figure 3.4 shows an overview of the experimental displays used in

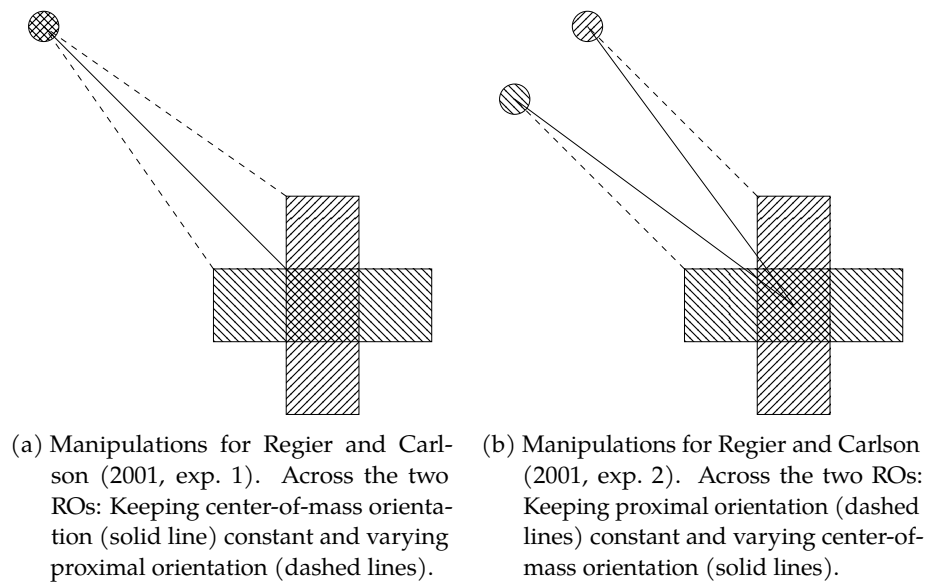


Figure 3.5: Examples of one LO placement for (a) exp. 1 and (b) exp. 2 from Regier and Carlson (2001). The two rectangular ROs that were used in the experiments are overlaid (filled with different patterns) to contrast the effect of the LO placement on the proximal (dashed lines) vs. center-of-mass orientation (solid lines). The fill pattern of the LO depicts with which RO the LO was shown.

these experiments. The aim of the third experiment was to explore the influence of the center-of-mass orientation in the region directly above the RO where the proximal orientation is constantly zero. Figure 3.9 depicts the two experimental displays for experiment 3, discussed in detail after the first two experiments.

EXPERIMENTS 1 AND 2 The way the LO positions were varied in the first two experiments is depicted in Figure 3.5 (for one sample LO). The figure shows both ROs at the same time (with different fill patterns) The LO is filled in the same style as the corresponding RO it was shown with. Center-of-mass orientations are depicted with solid lines, proximal orientations with dashed lines. In experiment 1, the LOs were placed at the same positions for both ROs and thus, the center-of-mass orientation for LOs (at the same grid-location) was kept constant across both ROs while the proximal orientation was different (see Figure 3.5a). In experiment 2, the LOs had different placements for both ROs (see Figure 3.5b). These placements allowed for the same proximal orientations (dashed lines in Figure 3.5b) while varying the center-of-mass orientations (solid lines).

For experiment 1, Regier and Carlson (2001) expected higher *above* ratings for the wide RO compared to the tall RO, because the proximal orientation is greater (i.e., it deviates more from canonical upright)

for the tall RO than for the wide RO. In contrast, they expected higher *above* ratings for the tall RO in experiment 2 compared to the wide RO, because here the center-of-mass orientation is greater (i.e., it deviates more from canonical upright) for the wide RO than for the tall RO. Indeed, they found this expected pattern.

Following the analyses presented in Regier and Carlson (2001), Table 3.2 shows the coefficients of linear regression models relating empirical and model-generated data. Crucially, to generate data with the cognitive models, I did not fit them to the experimental data from Regier and Carlson (2001) but used the parameters of the best model fits to the data from Logan and Sadler (1996, exp. 2, see Table 3.1). As can be seen from Table 3.2, all models account very well for the data (correlation always greater than 0.99) without using the data to estimate the best possible parameters. For comparison, I also provide the coefficients for the parameters that Regier and Carlson (2001) used (see rows “AVS (RC-LS fit)”). They also used the best parameters to fit the data from Logan and Sadler (1996, exp. 2), but their fit was not as close as my best fit (see Table 3.1). Nevertheless, using these parameters led to a slightly higher correlation compared to my best parameters as can be seen in Table 3.2. I will refer to the parameter set used by Regier and Carlson (2001) from now on as the RC-LS fit (“RegierCarlson-LoganSadler” fit).

Table 3.2: Linear model fits relating the empirical data from exps. 1 and 2 from Regier and Carlson (2001) with model-generated data for the same stimuli. I computed model-generated data with parameter values from the best fit to Logan and Sadler (1996, exp. 2, *above*) shown in Table 3.1 – except for lines denoted with “RC-LS fit” where I used parameter values from the AVS model fit to Logan and Sadler (1996, exp. 2, *above*) as reported in Regier and Carlson (2001).

Model	R ²	adj. R ²	y-intercept	slope	nRMSE
<u>Experiment 1, tall rectangle</u>					
AVS (RC-LS fit)	0.996	0.996	-0.614	1.088	0.054
AVS (my fit)	0.993	0.992	-0.615	1.073	0.059
rAVS _{prox}	0.994	0.994	-0.608	1.042	0.061
rAVS _{c-o-m}	0.992	0.991	-0.627	1.091	0.061
rAVS _{comb}	0.993	0.993	-0.623	1.064	0.061
rAVS _{w-comb}	0.992	0.991	-0.622	1.087	0.060
<u>Experiment 1, wide rectangle</u>					
AVS (RC-LS fit)	0.994	0.994	-0.323	1.060	0.040
AVS (my fit)	0.997	0.996	-0.348	1.057	0.036
rAVS _{prox}	0.995	0.994	-0.340	1.059	0.040
rAVS _{c-o-m}	0.997	0.997	-0.402	1.062	0.038
rAVS _{comb}	0.995	0.994	-0.480	1.072	0.048
rAVS _{w-comb}	0.997	0.997	-0.369	1.054	0.036

Table 3.2: Continued: Linear models for exps. 1 and 2 from Regier and Carlson (2001).

Model	R ²	adj. R ²	y-intercept	slope	nRMSE
Experiment 1, both ROs					
AVS (RC-LS fit)	0.994	0.994	-0.470	1.075	0.048
AVS (my fit)	0.994	0.994	-0.479	1.064	0.049
rAVS _{prox}	0.992	0.992	-0.473	1.050	0.052
rAVS _{c-o-m}	0.994	0.994	-0.512	1.076	0.051
rAVS _{comb}	0.993	0.993	-0.551	1.068	0.055
rAVS _{w-comb}	0.994	0.994	-0.493	1.070	0.050
Experiment 2, tall rectangle					
AVS (RC-LS fit)	0.993	0.993	-0.637	1.098	0.060
AVS (my fit)	0.991	0.991	-0.643	1.075	0.063
rAVS _{prox}	0.992	0.991	-0.633	1.046	0.066
rAVS _{c-o-m}	0.991	0.990	-0.657	1.096	0.064
rAVS _{comb}	0.991	0.991	-0.650	1.069	0.065
rAVS _{w-comb}	0.991	0.990	-0.652	1.092	0.064
Experiment 2, wide rectangle					
AVS (RC-LS fit)	0.995	0.994	-0.721	1.056	0.068
AVS (my fit)	0.996	0.996	-0.748	1.067	0.067
rAVS _{prox}	0.995	0.994	-0.740	1.066	0.067
rAVS _{c-o-m}	0.996	0.995	-0.802	1.069	0.072
rAVS _{comb}	0.992	0.992	-0.881	1.078	0.081
rAVS _{w-comb}	0.996	0.996	-0.766	1.061	0.069
Experiment 2, both ROs					
AVS (RC-LS fit)	0.992	0.992	-0.681	1.077	0.064
AVS (my fit)	0.993	0.993	-0.695	1.071	0.065
rAVS _{prox}	0.993	0.993	-0.684	1.056	0.067
rAVS _{c-o-m}	0.992	0.992	-0.730	1.082	0.068
rAVS _{comb}	0.991	0.991	-0.763	1.073	0.073
rAVS _{w-comb}	0.992	0.992	-0.710	1.077	0.066

EXP. 1, GOF According to Table 3.2, all models closely account for the data – without explicitly fitting the model parameters to the data. To investigate the goodness of the models in more detail, I also fitted all models directly to the data from experiments 1 and 2. The GOF values for experiment 1 are shown in Figure 3.6, separately for the tall and the wide rectangle. All models achieve low GOFs (lower than 0.041), i.e., all models closely fit the data. Evaluating models solely on their ability to closely fit data is not sufficient (Pitt & Myung, 2002; Roberts & Pashler, 2000) which is why I applied the SHO method (see Section 3.2.1).

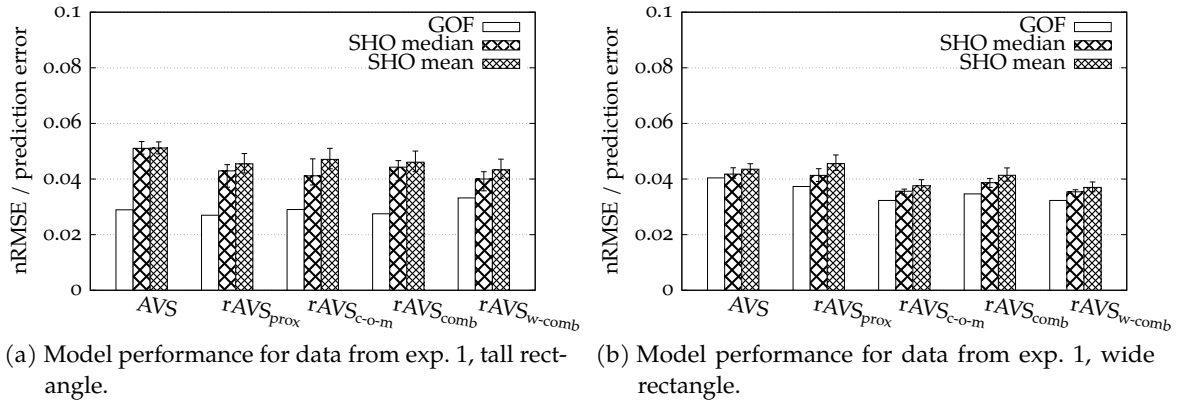


Figure 3.6: GOF and SHO results for fitting data from Regier and Carlson (2001, exp. 1): (a) tall rectangle, (b) wide rectangle. Error bars depict 95% confidence intervals of SHO median or mean respectively.

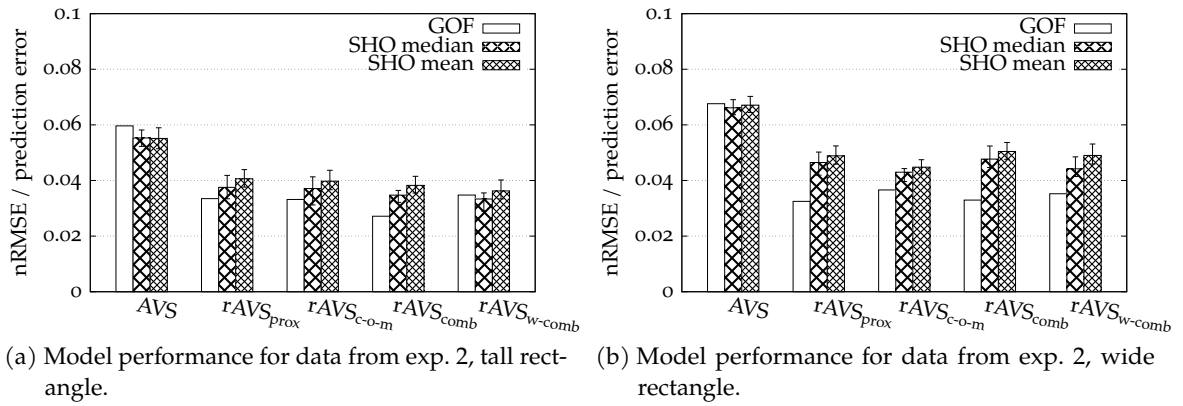


Figure 3.7: GOF and SHO results for fitting data from Regier and Carlson (2001, exp. 2): (a) tall rectangle, (b) wide rectangle. Error bars depict 95% confidence intervals of SHO median or mean respectively.

EXP. 1, SHO Figure 3.6 provides medians and means as outcome of the SHO method. For the tall rectangle (Figure 3.6a), all rAVS variations have quite similar SHO values, whereas the AVS model has a larger SHO value. The difference, however, is not large. Considering the wide rectangle (Figure 3.6b), the rAVSc-o-m and the rAVSw-comb models both have similar SHO values that are a bit lower than the SHO values of the other models.

EXP. 2, GOF & SHO The GOF values for the second experiment are displayed in Figure 3.7, separately for each RO. Again, all models fit the data from both ROs equally well, except for the AVS model which provides the worst GOF values for both ROs. However, this should not

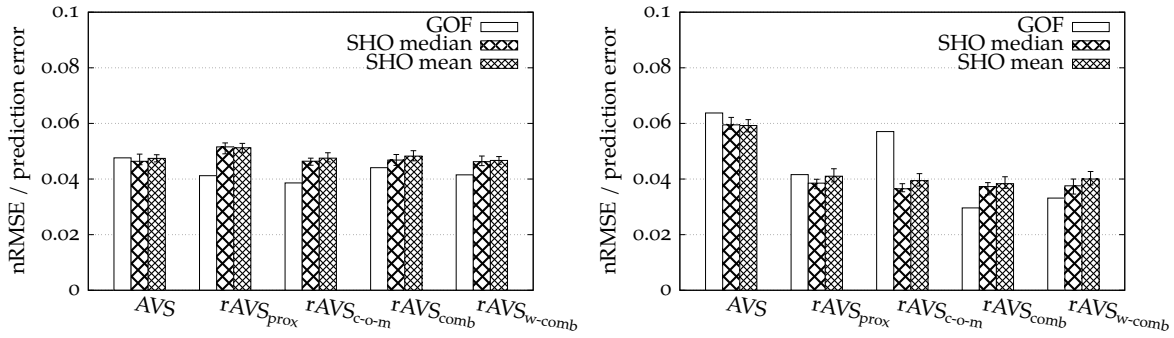


Figure 3.8: GOF and SHO results for fitting data from Regier and Carlson (2001, exps. 1 & 2, both ROs). Error bars depict 95% confidence intervals of SHO median or mean respectively.

be understood in the way that the AVS model cannot fit the data; its nRMSE is still very low (lower than 0.07).

Figure 3.7 also shows the medians and means from the SHO method. Considering these results, both ROs show a similar pattern. While all rAVS variations achieve similar low SHO values, the AVS model provides slightly worse (but still good) SHO values. It seems that these results disfavor the AVS model.

However, Regier and Carlson (2001) varied either one of center-of-mass orientation and proximal orientation throughout their first two experiments. They did so by either changing only the RO (experiment 1) or manipulating the locations of the LOs with respect to the used RO (experiment 2). Thus, the manipulation was always applied across both ROs. To capture the critical manipulations in the first experiment, the models must be fitted to the whole data set, i.e., the data from both ROs.

EXP. 1, BOTH ROS GOF values for data from both ROs of the first experiment are shown in Figure 3.8a. All models have similarly low GOF values. Looking at the SHO results also plotted in Figure 3.8a, the rAVS_{prox} model now performs slightly worse than any of the other models. The rest of the models perform almost equally well. This result is especially interesting, since the crucial manipulation in the first experiment was the variation of the proximal orientation (see Figure 3.5a). It seems that only using the center-of-mass orientation while ignoring the proximal orientation (as done by the rAVS_{c-o-m} model) results in a better model performance than the opposite (ignoring center-of-mass orientation but using the proximal orientation, as done by the rAVS_{prox} model). Based on the critical manipulation of experiment 1 (proximal orientation, not center-of-mass orientation), one would expect a contrary result here. Possibly the influence of the center-of-mass orien-

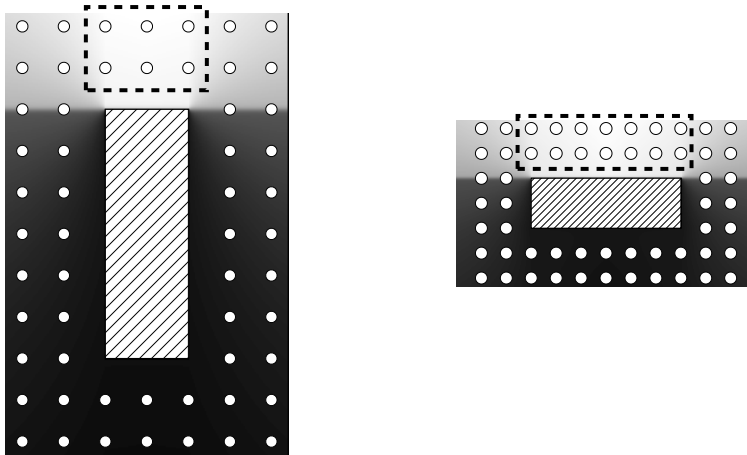


Figure 3.9: Displays used for simulating the stimuli of exp. 3 from Regier and Carlson (2001, reconstructed from data provided by Regier & Carlson, 2001). LOs are displayed as circles for visualization purposes. The simulations used single point LOs (at the center of the circles). Critical manipulation: In the regions directly above the ROs (marked with the dashed boxes), the proximal orientation is constant while the center-of-mass orientation varies. ROs are patterned for visualization purposes only. Backgrounds depict $rAVS_{w-comb}$'s spatial templates (lighter color coding higher rating) computed with best fitting parameters for data from Regier and Carlson (2001, exps. 1–3).

tation is stronger than the proximal orientation and thus shadowing the effect of the proximal orientation found by Regier and Carlson (2001, higher *above* ratings for the wide compared to the tall rectangle).

EXP. 2, BOTH ROS The GOF and SHO results for the data combined from both ROs for the second experiment can be found in Figure 3.8b. The AVS model and the $rAVS_{c-o-m}$ model have worse GOF values than the other models. The $rAVS_{comb}$ model has the best GOF. However, all GOFs are lower than 0.07, showing that all models fit the data well. Looking at the SHO results, the AVS model clearly obtains the worst value, while all $rAVS$ variations exhibit a similar performance.

EXP. 3 The third experiment conducted by Regier and Carlson (2001) also kept the proximal orientation constant while varying the center-of-mass orientation, comparable to the second experiment. In the second experiment, however, only two LOs were presented directly above the RO, which also shared the same proximal and center-of-mass orientation, whereas in the third experiment several different LO positions directly above the RO were tested. By doing so, the effect of the center-of-mass orientation in the region directly above the RO can be explored

Table 3.3: Linear model fits relating the empirical data from exp. 3 from Regier and Carlson (2001) with model-generated data for the same stimuli. I computed model-generated data with parameter values from the best fit to Logan and Sadler (1996, exp. 2, *above*) shown in Table 3.1 – except for lines denoted with “RC-LS fit” where I used parameter values from the AVS model fit to Logan and Sadler (1996, exp. 2, *above*) as reported in Regier and Carlson (2001).

Model	R ²	adj. R ²	y-intercept	slope	nRMSE
Experiment 3, tall rectangle					
AVS (RC-LS fit)	0.984	0.984	-0.596	1.060	0.070
AVS (my fit)	0.984	0.984	-0.585	1.018	0.075
rAVS _{prox}	0.980	0.980	-0.615	1.022	0.080
rAVS _{c-o-m}	0.980	0.979	-0.565	1.071	0.072
rAVS _{comb}	0.983	0.983	-0.590	1.044	0.072
rAVS _{w-comb}	0.981	0.980	-0.573	1.075	0.072
Experiment 3, wide rectangle					
AVS (RC-LS fit)	0.993	0.993	-0.407	1.017	0.052
AVS (my fit)	0.993	0.993	-0.398	1.054	0.046
rAVS _{prox}	0.990	0.990	-0.430	1.061	0.053
rAVS _{c-o-m}	0.989	0.989	-0.329	0.934	0.077
rAVS _{comb}	0.994	0.994	-0.380	0.996	0.054
rAVS _{w-comb}	0.991	0.990	-0.339	0.937	0.076

– while holding the proximal orientation constant for all these points. Figure 3.9 shows the experimental display used in the third experiment.

EXP. 3, ALL DATA POINTS All models are able to account for the data from experiment 3, as is evident from the high correlation coefficients in Table 3.3 (computed without fitting these data but the data from Logan & Sadler, 1996) as well as the GOF values displayed in Figure 3.10 (all lower than 0.07). The AVS model, however, has a rather large GOF for the tall rectangle (Figure 3.10a) that is also reflected in a worse SHO value compared to the other models. The rAVS_{prox} and the rAVS_{comb} models obtain slightly better SHO values than the other two rAVS variations for the tall rectangle. For the wide rectangle, all models provide similar SHO values, except for the rAVS_{c-o-m} model that has a worse SHO value (Figure 3.10b).

The third experiment was designed to test for the effects of center-of-mass orientation in the region directly above the RO while the proximal orientation was kept constant. It is surprising that the rAVS_{c-o-m} model obtains the worst SHO value for the wide rectangle (Figure 3.10b). Thus, to further explore the behavior of the models with respect to the critical placements of the LOs, the next section presents GOF and SHO

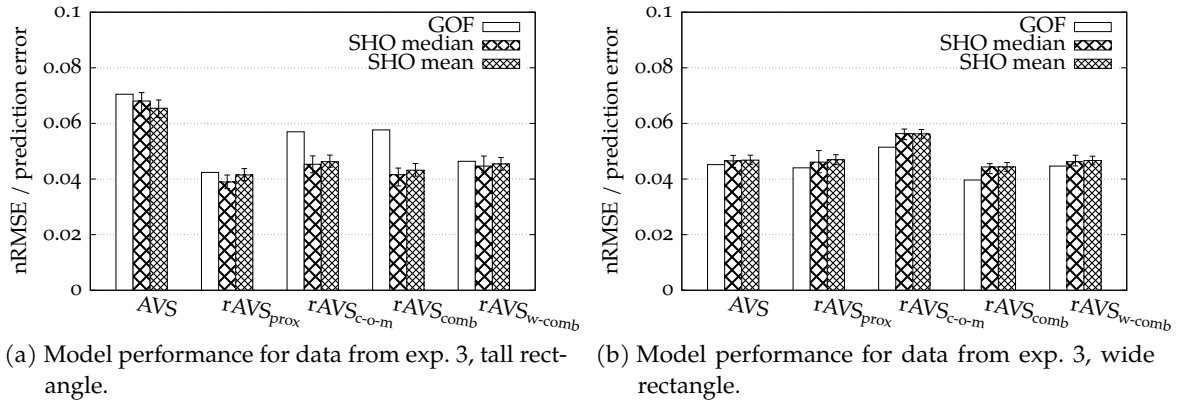


Figure 3.10: GOF and SHO results for fitting (Regier & Carlson, 2001, exp. 3). Error bars depict 95% confidence intervals of SHO median or mean respectively.

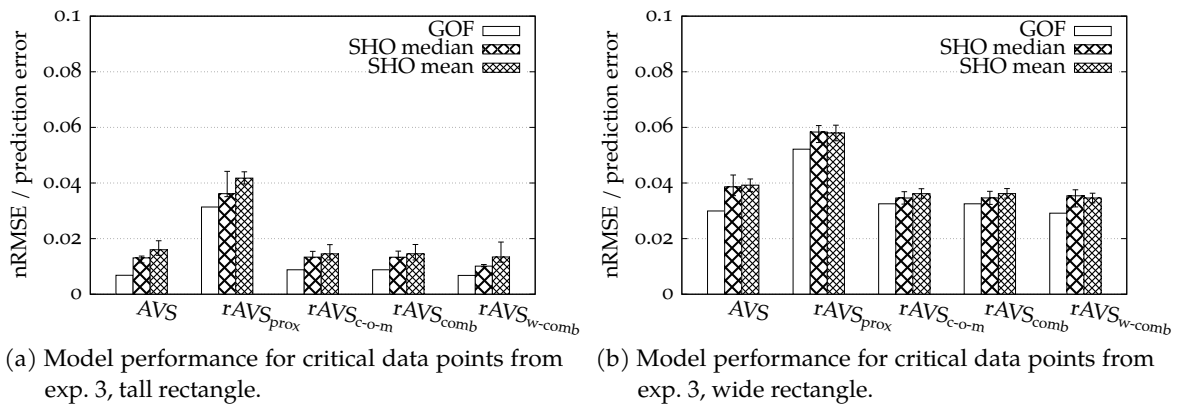


Figure 3.11: GOF and SHO results for fitting critical data points from Regier and Carlson (2001, exp. 3). Error bars depict 95% confidence intervals of SHO median or mean respectively.

results for data from these placements only. The critical LO positions are those that are depicted inside the dashed boxes in Figure 3.9.

EXP. 3, CRITICAL DATA POINTS Figure 3.11 shows the GOF values for the critical placements in experiment 3, separately for each RO. For both ROs, all models fit the data well (GOF smaller than 0.06; see Figure 3.11a and Figure 3.11b). The $rAVS_{prox}$ model, however, obtains a comparably bad fit. This disadvantage is confirmed by the SHO method: The $rAVS_{prox}$ model gets the worst SHO results for both ROs. All other models perform similarly, as indicated by very close medians and overlapping confidence intervals. Thus, the results for the critical LO placements provide evidence that the $rAVS_{prox}$ model does not

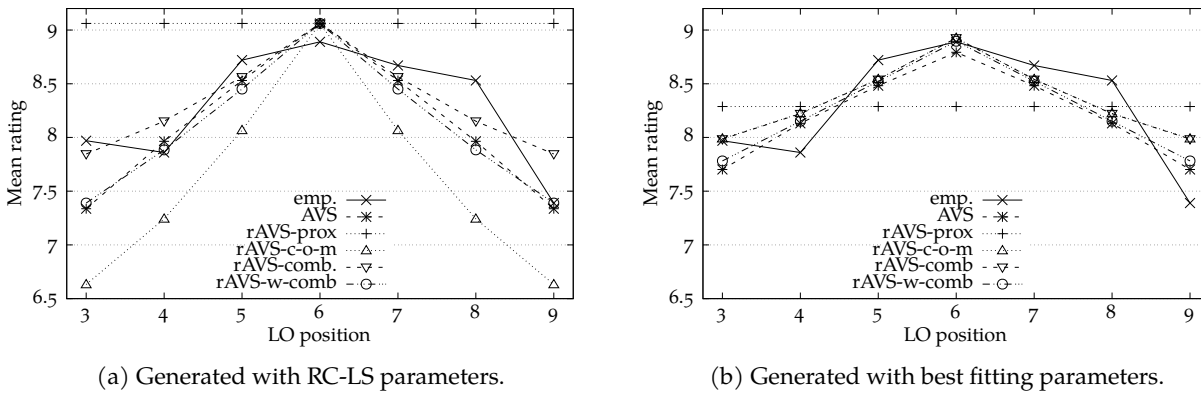


Figure 3.12: Qualitative comparison for Regier and Carlson (2001, exp. 3, wide rectangle, upper row of LOs) computed with (a) parameter values used by Regier and Carlson (2001) and (b) parameter values of my best fit to the critical data from the wide RO of exp. 3.

describe the data as well as the other models. Like Regier and Carlson (2001), I discuss next the qualitative behavior of the models to see if they can capture the qualitative empirical trends.

EXP. 3, QUALITATIVE FIT Figure 3.12 displays the empirical mean ratings as well as the model-generated ratings for the upper row of the critical LO placements from experiment 3, wide rectangle. Figure 3.12a shows ratings as generated with model parameters from the RC-LS fit (see Table 3.1 on page 48). For model-generated ratings in Figure 3.12b, I fitted the models to all critical placements of experiment 3. The empirical ratings shown in Figure 3.12 are peaking at the center (position 6) which is directly above the center-of-mass of the RO. Thus, the empirical ratings show an effect of center-of-mass orientation despite constant proximal orientation.

Clearly, the $rAVS_{prox}$ model is not able to capture this pattern in the empirical data. It gives the same rating throughout all positions. This is not surprising since the $rAVS_{prox}$ model only considers the proximal orientation which was kept constant by design. With these results however, the $rAVS_{prox}$ model can definitely be disqualified from the model competition. All other models approximate the data quite nicely. When fitting these other models to the data sets (results illustrated in Figure 3.12b), they generate almost indistinguishable rating patterns close to the empirical pattern.

The $rAVS_{prox}$ model is disconfirmed.

DISCUSSION EXPS. 1–3 Considering all results for the first three experiments, the $rAVS_{prox}$ model can be disqualified because it does not capture the empirical pattern from the third experiment (Figure 3.12). Interestingly, the AVS model performs worse (but still good) for the second experiment compared to the $rAVS$ variations (see Figure 3.8b).

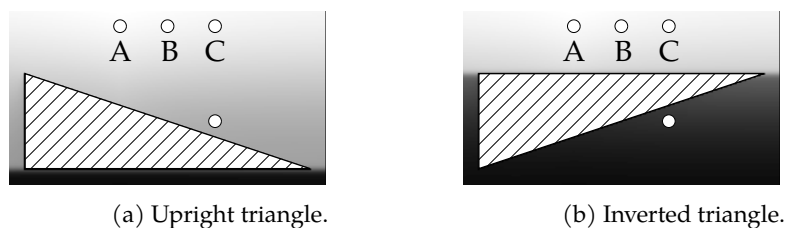


Figure 3.13: Displays used for simulating the stimuli of exp. 4 from Regier and Carlson (2001, reconstructed from data provided by Regier & Carlson, 2001). LOs are displayed as circles for visualization purposes. The simulations used single point LOs (at the center of the circles). Critical manipulation: “Point A is above the center of mass of the triangle, Point B is above the midpoint of the base of the triangle, and Point C is placed so that its distance from B equals the distance between A and B” (Regier & Carlson, 2001, p. 285). ROs are patterned for visualization purposes only. Backgrounds depict $rAVS_{w-comb}$'s spatial templates (lighter color coding higher rating) computed with best fitting parameters for data from Regier and Carlson (2001, exp. 4).

The performance of all other models, however, cannot be distinguished from each other. This is interesting because the $rAVS$ variations that used either only the center-of-mass orientation or only the proximal orientation (i.e., $rAVS_{c-o-m}$ and $rAVS_{prox}$) did not perform better in any of the first two experiments although these experiments explicitly tested for the center-of-mass orientation and the proximal orientation.

3.2.4 Dissociate Center-of-Mass from Midpoint: Regier and Carlson (2001, Exp. 4)

In the third experiment, the LO position at which the ratings peaked was directly above the center-of-mass but also directly above the midpoint of the base of the RO. Experiment 4 was designed to dissociate the center-of-mass from the midpoint of the RO. To this end, Regier and Carlson (2001) used two triangles that are depicted in Figure 3.13. Here, the center-of-mass is at a different point than the midpoint of the base of the triangle. The critical LO positions were the three points A, B, and C above the triangle: “Point A is above the center of mass of the triangle, Point B is above the midpoint of the base of the triangle, and Point C is placed so that its distance from B equals the distance between A and B” (Regier & Carlson, 2001, p. 285). For the upright triangle (Figure 3.13a), the empirical mean rating was significantly lower for point C than for points A and B. For the inverted triangle (Figure 3.13b), the mean rating for point C was significantly lower than the mean rating for point B but it was not significantly lower than the mean rating for

Table 3.4: Linear model fits relating the empirical data from exp. 4 from Regier and Carlson (2001) with model-generated data for the same stimuli. I computed model-generated data with parameter values from the best fit to Logan and Sadler (1996, exp. 2, *above*) shown in Table 3.1 – except for lines denoted with “RC-LS fit” where I used parameter values from the AVS model fit to Logan and Sadler (1996, exp. 2, *above*) as reported in Regier and Carlson (2001).

Model	R ²	adj. R ²	y-intercept	slope	nRMSE
Experiment 4, both ROs					
AVS (RC-LS fit)	0.988	0.985	-1.050	1.164	0.057
AVS (my fit)	0.993	0.992	-1.063	1.187	0.060
rAVS _{prox}	0.991	0.989	-1.045	1.192	0.066
rAVS _{c-o-m}	0.933	0.922	-1.498	1.138	0.112
rAVS _{comb}	0.980	0.976	-1.272	1.164	0.067
rAVS _{w-comb}	0.952	0.944	-1.327	1.124	0.095

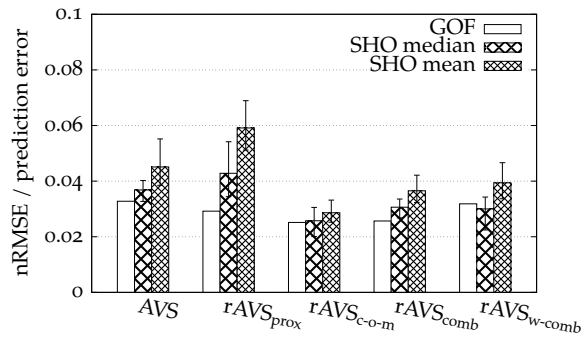
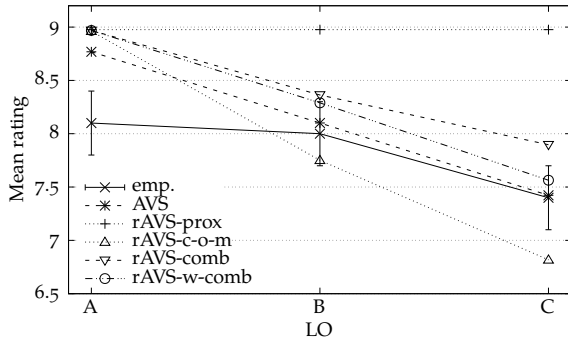


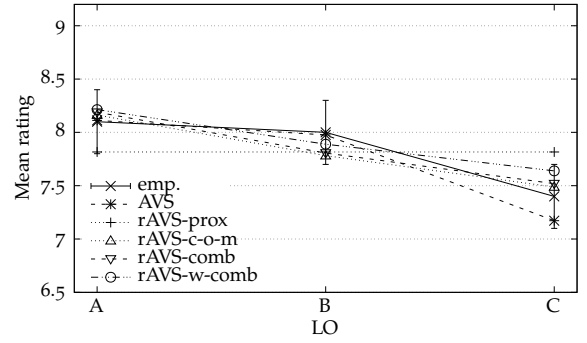
Figure 3.14: GOF and SHO results for fitting Regier and Carlson (2001, exp. 4, both ROs). Error bars depict 95% confidence intervals of SHO median or mean respectively.

point A. For both ROs, the mean ratings for point A and point B did not differ significantly (see also Figure 3.15 for a visualization of the empirical findings).

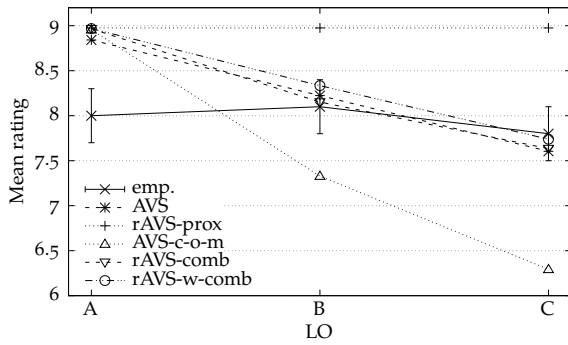
The correlation for empirical and model-generated data for the fourth experiment presented in Table 3.4 is very high. Again, all models closely account for the data without being fit to it (remember that I used model parameters for fitting the data from Logan & Sadler, 1996, exp. 2, *above*, for these tables). The following GOF and SHO results must be considered carefully, because the data set from experiment 4 consists of only four data points for each RO. This is especially crucial for the SHO method, which splits the data in a training set of only three data points and a test set of only one data point. This is why I have combined the data from both ROs to one data set consisting of 8 data points. The results for these data are presented in Figure 3.14.



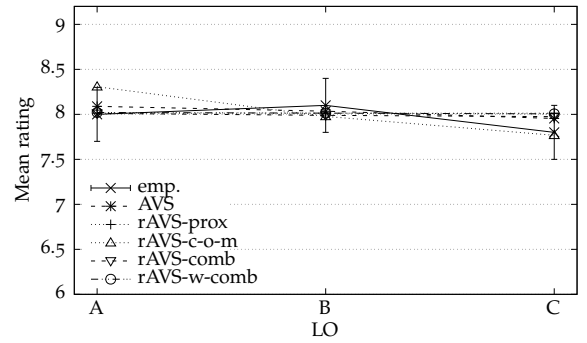
(a) Upright triangle, generated with RC-LS parameters.



(b) Upright triangle, generated with best fitting parameters.



(c) Inverted triangle, generated with RC-LS parameters.



(d) Inverted triangle, generated with best fitting parameters.

Figure 3.15: Qualitative comparison for data from Regier and Carlson (2001, exp. 4): (a, b) upright triangle, (c, d) inverted triangle. Computed with (a, c) parameter values used by Regier and Carlson (2001) and (b, d) parameter values of my best fit to the corresponding data. Error bars for the empirical data depict the ± 0.3 difference needed for significance (based on 95% confidence intervals) as reported by Regier and Carlson (2001, p. 285).

Considering the GOF results, all models closely fit the data for both ROs: all GOFs are lower than 0.04. The $rAVS_{c-o-m}$ model provides the best SHO value although the confidence intervals overlap with the $rAVS_{comb}$ model. The $rAVS_{prox}$ model has the worst SHO value, while the AVS and the $rAVS_{w-comb}$ model provide in-between results. However, since the SHO results might be flawed due to the small number of available data points, it is even more important to look at the qualitative behavior of the models.

QUALITATIVE FIT Figure 3.15 shows the empirical data and the model-generated data for the three critical LO positions, separately for each RO. Figures 3.15a and 3.15c are displaying model-generated ratings computed with the RC-LS parameters. Figures 3.15b and 3.15d show model-generated ratings fitted to all four data points from the corre-

sponding RO. Furthermore, the difference that was needed for two ratings to be significantly different is plotted as error bars in Figure 3.15. This difference is 0.3 and is based on 95% confidence intervals, as reported by Regier and Carlson (2001, p. 285).

The already disconfirmed $rAVS_{prox}$ model again is not able to capture the trend that can be seen in the empirical data. Again, this finding is not surprising: The proximal orientation is the same for all three LOs A, B, and C and thus, the rating that $rAVS_{prox}$ generates does not differ between the positions. Accordingly, these qualitative results once more disqualify the $rAVS_{prox}$ model.

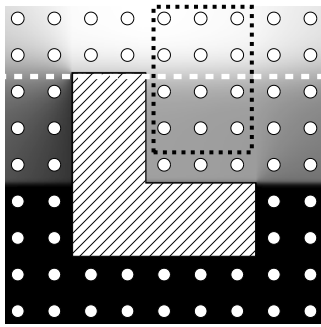
The other models capture the empirical data quite well. For the upright triangle, all models generate very similar data close to the empirical data: all artificial ratings differ less from the corresponding empirical ratings than the difference needed for significance (± 0.3 ; Figure 3.15b). When fitted to the data from the inverted triangle (Figure 3.15d), the models show almost the same behavior as the $rAVS_{prox}$ model: The same rating for every position (with the $rAVS_{c-o-m}$ model providing the greatest exception from this straight line). This comes as no surprise because the empirical data points are also very close to each other: 8.0 (A), 8.1 (B), 7.8 (C). Using the parameters fitted on the Logan and Sadler (1996) data, the models show higher ratings for A than for C (Figure 3.15a). Note, however, that Regier and Carlson (2001) did not find significantly different ratings for point A and point C for the inverted triangle (only the difference between ratings for point C and point B was significant).⁶ My empirical study with asymmetrical ROs provides more pertinent data suggesting that for ROs with a flat top (such as the inverted triangle), people consider the center-of-object more than the center-of-mass (Section 4.2.1).

DISCUSSION EXP. 4 Taken together, the low number of LOs in the fourth experiment makes it difficult to use these data for quantitative model assessment with the SHO method. Considering the SHO results, the $rAVS_{prox}$ performs slightly worse than the other models. Indeed, the comparably bad SHO value of the $rAVS_{prox}$ model is confirmed by its bad qualitative behavior. This disqualifies the $rAVS_{prox}$ model once more. All other models are able to reproduce the qualitative behavior. Accordingly, none of the other models can be disfavored.

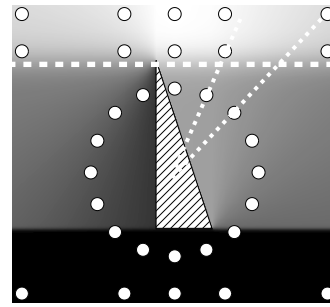
3.2.5 *The Effect of the Grazing Line: Regier and Carlson (2001, Exps. 5 & 6)*

Experiments 5 and 6 tested for the effect of the 'grazing line' on the acceptability of spatial terms. The grazing line is an imaginary horizon-

⁶ See also Lovett and Forbus (2009) who failed to replicate this effect with their computational model, too, and also point to the small effect size and the low number of data points.



(a) Experiment 5. Black dotted box frames six critical LOs above and six critical LOs below the grazing line.



(b) Experiment 6. Critical LOs are two pairs of LOs that share both the same proximal & center-of-mass orientations but are on different sides of the grazing line. Each pair is connected with a dotted line (center-of-mass orientation).

Figure 3.16: Displays used for simulating the stimuli of exps. 5 & 6 from Regier and Carlson (2001, reconstructed from data provided by Regier & Carlson, 2001). LOs are displayed as circles for visualization purposes. The simulations used single point LOs (at the center of the circles). Critical manipulation: LOs placed above or below the grazing line (horizontal white dashed line). ROs are patterned for visualization purposes only. Backgrounds depict $rAVS_{w-comb}$'s spatial templates (lighter color coding higher rating) computed with best fitting parameters for data from Regier and Carlson (2001, exps. 5 & 6).

tal line that touches the top-most point of the RO. The stimuli used in experiments 5 and 6 and the grazing line are depicted in Figure 3.16. In both experiments, Regier and Carlson (2001) found an influence of the grazing line: Points that are above the grazing line are rated higher than points that are below the grazing line. This remains true even if the center-of-mass and proximal orientations were held constant (see experiment 6).

EXP. 5 The RO and all LO positions for the fifth experiment are depicted in Figure 3.16a. Critical LOs are the twelve points inside the dashed box, from which six are above the grazing line and six are below the grazing line. Regier and Carlson (2001) found that the six LOs above the grazing line were rated higher than the six LOs below the grazing line. As a measure of the strength of this effect, they subtracted the average rating for the lower LOs from the average rating for the upper LOs. This measure is positive if the average rating for the upper LOs is higher than the average rating for the lower LOs. Like Regier and Carlson (2001), I also computed this measure for the models, using the parameters for fitting the data from Logan and Sadler (1996). The results can be found in Table 3.5. All models show a grazing line effect with a similar strength as observed empirically.

Table 3.6 shows coefficients of linear model fits for the empirical data from experiment 5 and model-generated data using the parameters for

Table 3.5: The effect of the grazing line for exps. 5 and 6 using parameters for fitting Logan and Sadler (1996, exp. 2, *above*, Table 3.1).

Model	exp. 5	exp. 6	
		left pair	right pair
empirical	3.528	1.730	3.885
AVS (RC-LS)	3.736	3.038	3.268
AVS (my fit)	3.952	3.549	3.339
rAVS _{prox}	4.042	3.579	3.179
rAVS _{c-o-m}	4.142	3.659	3.331
rAVS _{comb}	4.177	3.892	3.246
rAVS _{w-comb}	3.904	3.541	3.318

Table 3.6: Linear model fits relating the empirical data from exps. 5 and 6 from Regier and Carlson (2001) with model-generated data for the same stimuli. I computed model-generated data with parameter values from the best fit to Logan and Sadler (1996, exp. 2, *above*) shown in Table 3.1 – except for lines denoted with “RC-LS fit” where I used parameter values from the AVS model fit to Logan and Sadler (1996, exp. 2, *above*) as reported in Regier and Carlson (2001).

Model	R ²	adj. R ²	y-intercept	slope	nRMSE
Experiment 5					
AVS (RC-LS fit)	0.976	0.976	-0.461	0.944	0.094
AVS(my fit)	0.978	0.978	-0.526	0.972	0.088
rAVS _{prox}	0.969	0.969	-0.614	0.995	0.095
rAVS _{c-o-m}	0.959	0.958	-0.450	0.918	0.113
rAVS _{comb}	0.976	0.976	-0.547	0.959	0.096
rAVS _{w-comb}	0.974	0.973	-0.542	0.940	0.104
Experiment 6					
AVS (RC-LS fit)	0.928	0.926	0.179	1.167	0.122
AVS (my fit)	0.928	0.926	0.142	1.153	0.115
rAVS _{prox}	0.910	0.907	0.048	1.152	0.120
rAVS _{c-o-m}	0.906	0.903	0.323	1.174	0.144
rAVS _{comb}	0.932	0.930	0.164	1.180	0.124
rAVS _{w-comb}	0.918	0.915	0.273	1.170	0.134

fitting the data from Logan and Sadler (1996). All correlations are high, providing evidence that all models closely fit the data. This is confirmed by the good GOFs for all models, presented in Figure 3.17. Considering all LOs, all nRMSEs are lower than 0.09 (Figure 3.17a). Fitting only the critical data points (six LOs above and six LOs below the grazing line), the nRMSE is even lower than 0.04 for all models (Figure 3.17b).

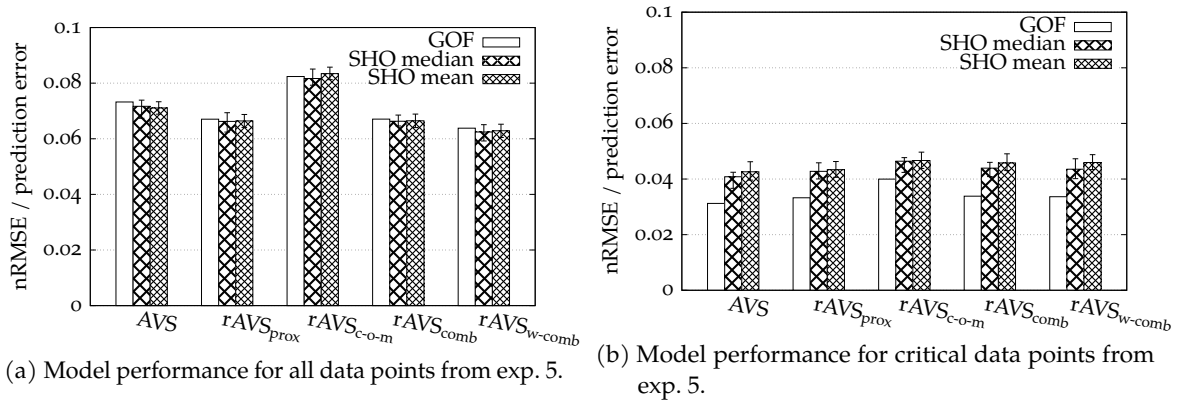


Figure 3.17: GOF and SHO results for fitting Regier and Carlson (2001, exp. 5). Error bars depict 95% confidence intervals of SHO median or mean respectively.

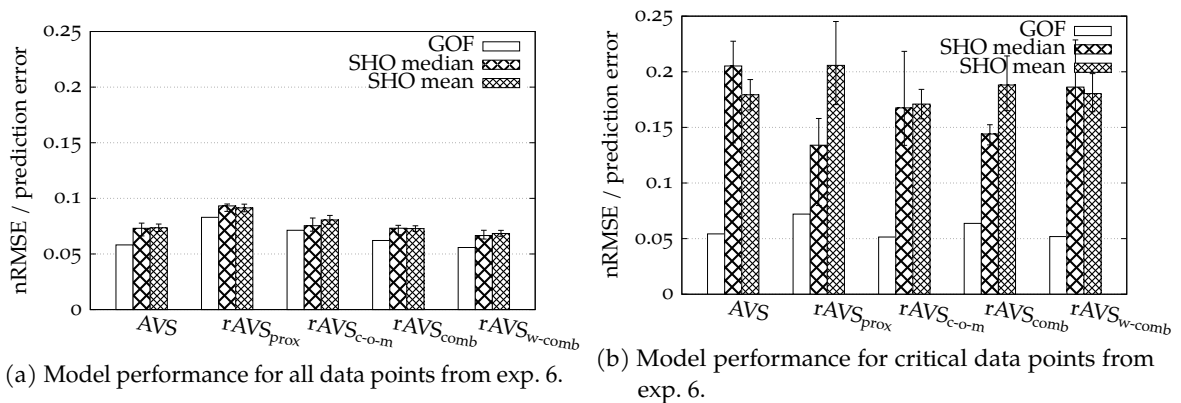


Figure 3.18: GOF and SHO results for fitting Regier and Carlson (2001, exp. 6). Error bars depict 95% confidence intervals of SHO median or mean respectively.

Looking at the SHO results, all models are indistinguishable for the critical data set (Figure 3.17b). Using the full data set (Figure 3.17a), the rAVS_{w-comb} model obtains a slightly better SHO value than all other models, while the AVS model has a slightly worse SHO value than all models – except for the rAVS_{c-o-m} model which gets the worst SHO value.

Taken together, the results for the fifth experiment slightly favor the rAVS_{w-comb} model and disfavor the rAVS_{c-o-m} model. However, these results only emerge for the whole data set and not for the critical subset (six LOs above and six LOs below the grazing line). Furthermore, note that all of the models show a comparable performance regarding the strength of the effect of the grazing line, shown in Table 3.5. Thus, qualitatively all models seem to be able to accommodate the results.

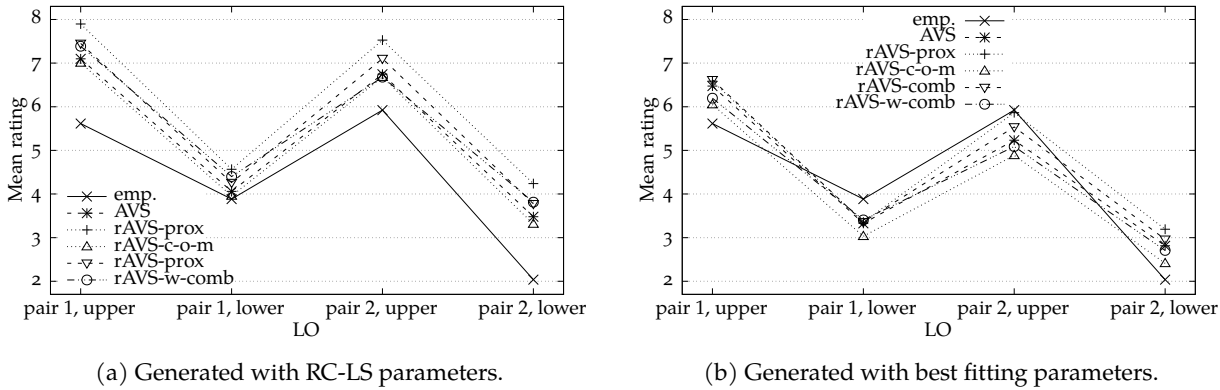


Figure 3.19: Qualitative comparison for Regier and Carlson (2001, exp. 6, critical LOs) computed with (a) parameter values used by Regier and Carlson (2001) and (b) parameter values of my best fit to all data from exp. 6.

EXP. 6 The sixth experiment again tested for the effect of the grazing line but this time the proximal and the center-of-mass orientation were also controlled. The used experimental display is shown in Figure 3.16b. There were two pairs of critical LOs (connected with dotted lines in Figure 3.16b) that shared the same center-of-mass orientation and proximal orientation (in one pair), but one LO of each pair was placed above the grazing line and the other one below the grazing line. Thus, any different ratings for these LOs cannot be explained by differences in the center-of-mass or proximal orientations, but by their locations relative to the grazing line.

As in experiment 5, Regier and Carlson (2001) expected and found higher ratings for LOs above the grazing line compared to LOs below the grazing line. Table 3.5 shows the strength of this effect for each model using the same measure as used in the fifth experiment (rating for the upper LO subtracted by the rating for the lower LO) – computed with parameters from fitting data from Logan and Sadler (1996, exp. 2, *above*). Again, all models qualitatively replicate the grazing line effect but the replication is quantitatively not as close as for experiment 5. Using the same model parameters, Table 3.6 prints the correlation of the model-generated data to the empirical data. The overall correlation is worse than for the previous experiments (but still higher than 0.9), with the correlation for the $rAVS_{c-o-m}$ model being the lowest.

Nonetheless, all models fit the data well, as is evident from the GOFs plotted in Figure 3.18a: All GOFs are lower than 0.09. Considering the SHO results, the $rAVS_{prox}$ model gets a slightly worse result than all other models for all LO positions (Figure 3.18a). All other models cannot be distinguished on these data. The SHO results for the critical subset shown in Figure 3.18b are not very reliable, since the critical subset consists of only four data points. Hence, the SHO medians,

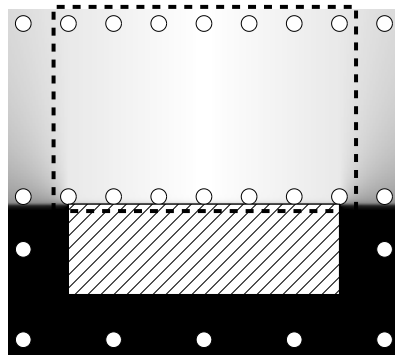


Figure 3.20: Displays used for simulating the stimuli of exp. 7 from Regier and Carlson (2001, reconstructed from data provided by Regier & Carlson, 2001). LOs are displayed as circles for visualization purposes. The simulations used single point LOs (at the center of the circles). Critical manipulation: distance of LOs to RO. The dashed box frames the critical points. RO is patterned for visualization purposes only. Background depicts $rAVS_{w-comb}$'s spatial templates (lighter color coding higher rating) computed with best fitting parameters for data from Regier and Carlson (2001, exp. 7).

means and their corresponding confidence intervals are relatively large and cannot be taken as support for either model. Accordingly, I discuss the qualitative fit to the data that is depicted in Figure 3.19.

Again, Figure 3.19a shows the model output with the parameters used in Regier and Carlson (2001, "RC-LS fit") and Figure 3.19b shows the model output with the parameters that gave the closest fit to all data from experiment 6. As is evident from both figures, all models accommodate the trend in the empirical data. Based on this qualitative comparison, no model can be disqualified for the four critical points of experiment 6. Considering all data points for experiment 6, however, the $rAVS_{prox}$ model still performs worst in terms of GOF and SHO (Figure 3.18a), which further supports its disqualification.

3.2.6 *The Effect of Distance: Regier and Carlson (2001, Exp. 7)*

The last experiment in Regier and Carlson (2001) was designed to test for the effect of distance between the RO and the LO on the acceptability of spatial prepositions. To this end, Regier and Carlson (2001) used a wide rectangle as RO and placed LOs at two distances above the RO (see Figure 3.20). What Regier and Carlson (2001) expected and found was that the upper row of LOs is more "sensitive to the centeredness of the trajector [LO] above the landmark [RO]" (Regier & Carlson, 2001, p. 289) than the lower row of LOs. Indeed, the ratings for the upper

Table 3.7: Linear model fits relating the empirical data from exp. 7 from Regier and Carlson (2001) with model-generated data for the same stimuli. I computed model-generated data with parameter values from the best fit to Logan and Sadler (1996, exp. 2, *above*) shown in Table 3.1 – except for lines denoted with “RC-LS fit” where I used parameter values from the AVS model fit to Logan and Sadler (1996, exp. 2, *above*) as reported in Regier and Carlson (2001).

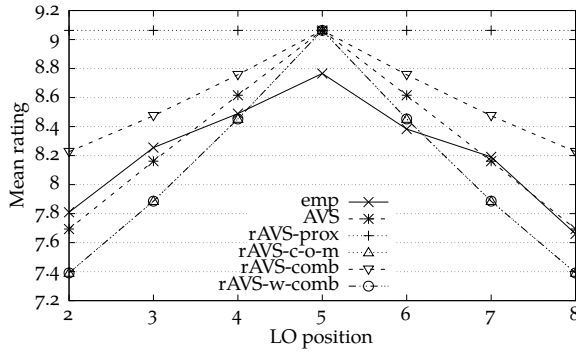
Model	R ²	adj. R ²	y-intercept	slope	nRMSE
Experiment 7					
AVS (RC-LS fit)	0.965	0.963	-0.836	1.068	0.090
AVS (my fit)	0.965	0.963	-0.755	1.076	0.085
rAVS _{prox}	0.955	0.953	-0.807	1.087	0.095
rAVS _{c-o-m}	0.903	0.899	-0.612	0.960	0.145
rAVS _{comb}	0.963	0.961	-0.726	1.071	0.086
rAVS _{w-comb}	0.937	0.934	-0.675	0.984	0.124

row show a greater peak for the central positions, whereas the ratings for the lower row are almost flat (see empirical data in Figure 3.21).

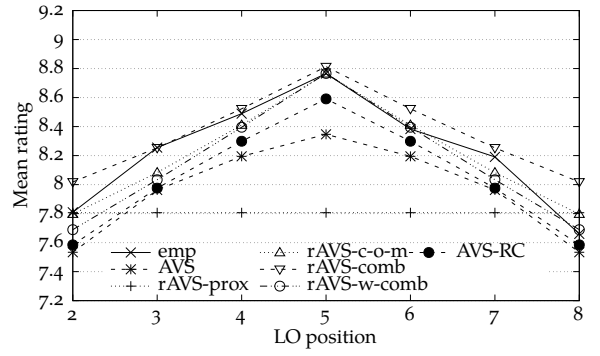
This pattern of result is also predicted by the AVS model. The rAVS_{prox} model does not predict such a rating pattern: It generates equal ratings for LOs in one row. The rAVS_{c-o-m} model, however, predicts the opposite than the AVS model and thus also conflicts with the empirical findings: LOs close to the RO have a higher center-of-mass orientation than more distant LOs (keeping the horizontal component constant). Thus, the rAVS_{c-o-m} model predicts more sensitivity to the centeredness for the lower row instead of for the upper row. Consequently, the rAVS_{c-o-m} model-generated data (with best fitting parameters for Logan & Sadler, 1996) provide the worst correlation to the data from experiment 7 (though still high; see Table 3.7) – compared to the data from the other models.

QUALITATIVE COMPARISON Figure 3.21 shows the output of the models for the critical points in experiment 7, separately for each row and computed with the parameters of the RC-LS fit (Figures 3.21a and 3.21c) or the parameters that gave the closest fit to all data from experiment 7 (Figures 3.21b and 3.21d). Note that Regier and Carlson (2001) also fitted the whole data set from experiment 7 with the AVS model and provided the model parameters of their fit. On the critical subset, the parameters of their fit result in a similar nRMSE compared to mine but a better correlation ($R^2: 0.89 > 0.73$). I generated data with parameters from both fits in Figures 3.21b and 3.21d with the fit reported by Regier and Carlson (2001) labeled as AVS-RC.

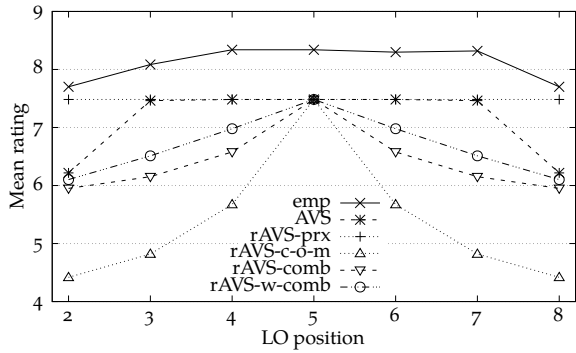
As discussed, the empirical ratings for the upper row (Figures 3.21a and 3.21b) peak in the middle, whereas the empirical ratings for the



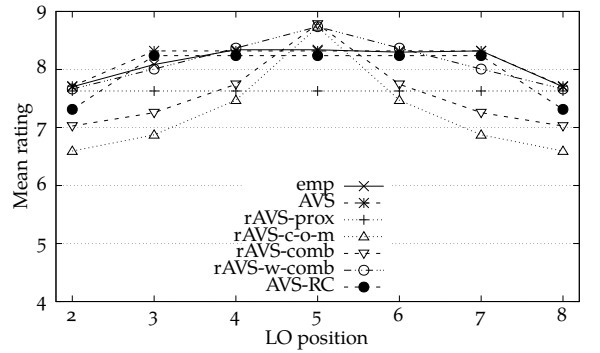
(a) Upper row, generated with RC-LS parameters.



(b) Upper row, generated with best fitting parameters.



(c) Lower row, generated with RC-LS parameters.



(d) Lower row, generated with best fitting parameters.

Figure 3.21: Qualitative comparison for Regier and Carlson (2001, exp. 7): (a, b) upper row of LOs, (c, d) lower row of LOs. Generated with (a, c) parameter values used by Regier and Carlson (2001) and (b, d) parameter values of my best fit to all data from exp. 7.

lower row (Figure 3.21c and 3.21d) show a flat profile for the four central LOs. The $rAVS_{prox}$ model gives the same rating for all LOs with the same elevation⁷ and thus cannot accommodate the different empirical trends in both rows. All other models show a peak for the upper row (Figure 3.21a and 3.21b). For the lower row, however, both the $rAVS_{c-o-m}$ model and the $rAVS_{comb}$ model cannot capture the flat rating profile from the empirical trend (Figure 3.21d). The AVS model generates data with a flat profile very close to the empirical data. Although the $rAVS_{w-comb}$ model-generated data do not show a completely flat rating profile, they closely fit the data.

GOF AND SHO The wrong prediction of the $rAVS_{c-o-m}$ model is reflected in its GOF values for all data points (Figure 3.22a). Here, the GOF value of the $rAVS_{c-o-m}$ model is worse than that of all other models. For the critical subset, however, all models obtain very good GOF results

⁷ The difference between the rows despite equal proximal orientation originates from the height component.

(all lower than 0.04, see Figure 3.22b). Looking at the SHO results, the $rAVS_{c-o-m}$ model gets the worst result for all data points (Figure 3.22a). Here, the AVS model, the $rAVS_{prox}$ model and the $rAVS_{w-comb}$ model perform almost similar. The $rAVS_{comb}$ model provides a slightly worse SHO result. For the critical subset, the AVS model has the best SHO result but the $rAVS_{w-comb}$ model is only slightly worse (Figure 3.22b). The other models have worse SHO results.

THE SAME RO: EXPERIMENT 3 AND 7 Since the RO used in experiment 7 was the same that was already used in experiment 3 (wide rectangle), I have also fitted the models to the combined data from both experiments. Using the data from both experiments provides a more complete empirical spatial template – instead of only two rows of LOs above the RO in each experiment, the combined data set consists of ratings for four rows of LOs above the RO (cf. Figure 3.9 on page 55 and Figure 3.20). In Figure 3.23, the GOF and SHO results for the combined data set are plotted. For Figure 3.23a, I have used all LO placements; for Figure 3.23b, I have only used the critical placements for each experiment.

The pattern of the results for the combined data set (Figure 3.23a) are very similar to the pattern for data from experiment 7 only (Figure 3.22a). Besides overall lower GOF and SHO values, the main differences in relative model performances are: The $rAVS_{prox}$ model obtains better results and the $rAVS_{comb}$ model provides worse results. Note, however, that the performance of the $rAVS_{prox}$ model on previous data already disqualified it. The worse performance of the $rAVS_{comb}$ model probably corresponds to its failure to accommodate the qualitative pattern from experiment 7. The results for the critical positions from the combined data set (Figure 3.23b) are also similar to the results for the critical data from experiment 7 only (Figure 3.22b). For the combined data, however, the $rAVS_{w-comb}$ model obtains slightly better results than the AVS model.

DISCUSSION EXP. 7 Both qualitative and quantitative simulation results for experiment 7 disconfirm the $rAVS_{c-o-m}$ and the $rAVS_{comb}$ models. In particular, the $rAVS_{c-o-m}$ model makes contradicting qualitative predictions compared to the empirical findings. The prediction from the AVS model is in line with the empirical evidence. This prediction stems from the attentional distribution implemented in the AVS model. At lower elevations, the attentional beam is smaller and thus not the whole RO is considered. This leads to a smaller effect of centeredness compared to higher elevations where the attentional beam is bigger and thus the RO gets more completely accounted for (cf. Regier & Carlson, 2001, p. 279).

The attentional distribution, however, is not of importance for the $rAVS$ models when only single-point LOs are used (see page 34). With

The $rAVS_{c-o-m}$ and the $rAVS_{comb}$ models are disconfirmed.

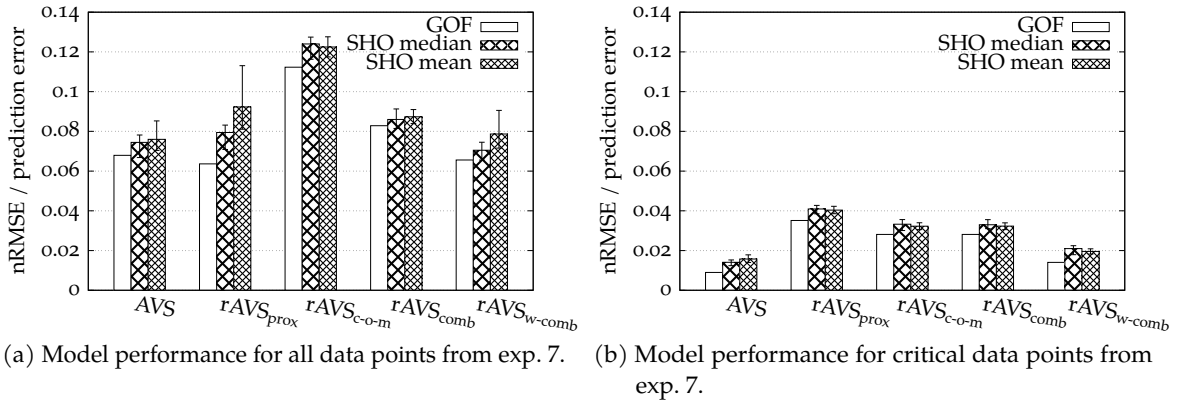


Figure 3.22: GOF and SHO results for fitting Regier and Carlson (2001, exp. 7). Error bars depict 95% confidence intervals of SHO median or mean respectively.

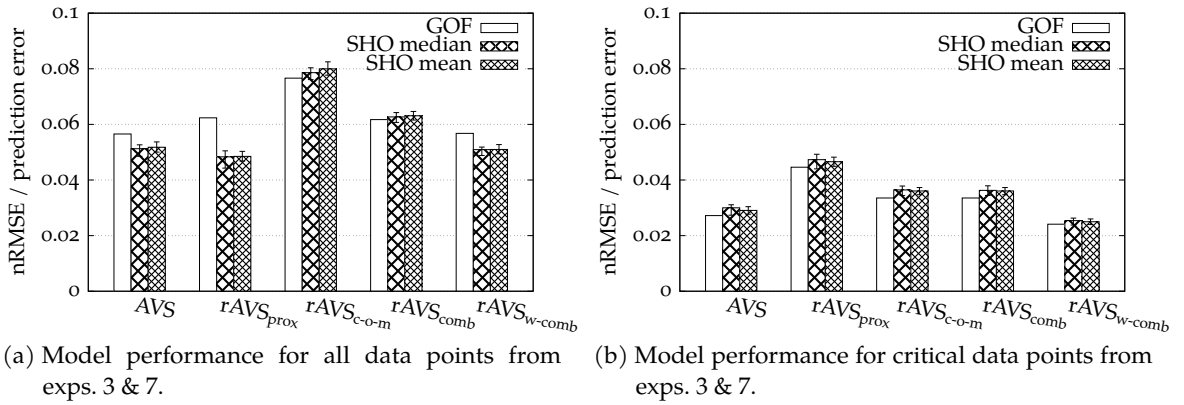


Figure 3.23: GOF and SHO results for fitting Regier and Carlson (2001, exps. 3 & 7, same RO). Error bars depict 95% confidence intervals of SHO median or mean respectively.

such simplified LOs, the $rAVS$ models compute only one vector for each rating. Since only the deviation but not the length of this single vector is important, it does not matter with how much attention this single vector is weighted (if the amount of attention is not zero). Thus, the attentional distribution does not change the way the $rAVS$ models are generating their output for single-point LOs.

In fact, this last experiment motivated me to develop the $rAVS_{w-comb}$ model. This $rAVS$ variation has a mechanism that weights the importance of the proximal orientation and the center-of-mass orientation according to the relative distance of the LO to the RO. As can be seen in the results plot in the previous sections, the $rAVS_{w-comb}$ model obtains very good results throughout all experiments. In particular, it performs comparable to the AVS model for experiment 7. As we will see in the next section, the $rAVS_{w-comb}$ model also works well for the combined

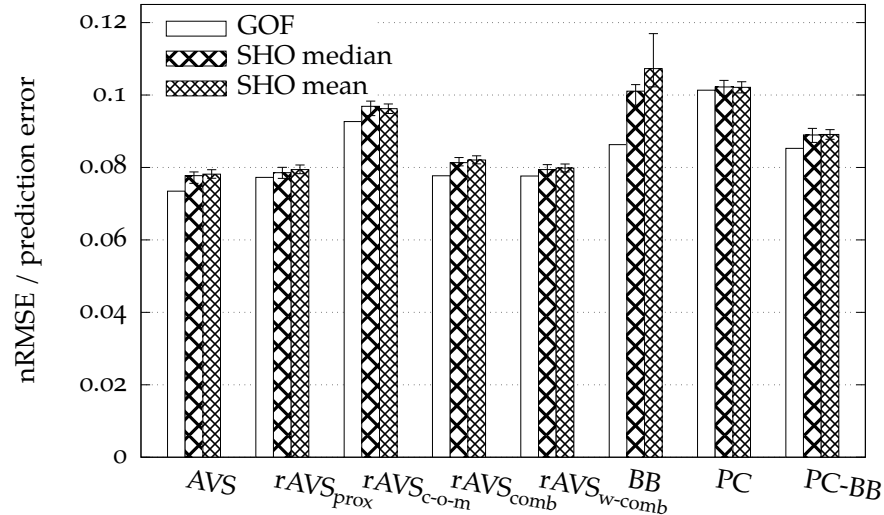


Figure 3.24: GOF and SHO results for fitting the data from all experiments from Regier and Carlson (2001). Error bars depict 95% confidence intervals of SHO median or mean respectively.

The success of the $rAVS_{w-comb}$ model is based on its relative distance mechanism.

data from all experiments. Thus, weighting the importance of center-of-mass orientation and proximal orientation via relative distance as proposed by the $rAVS_{w-comb}$ model provides an equally well performing mechanism compared to the attentional distribution of the AVS model.

3.2.7 All Experiments from Regier and Carlson (2001)

In Figure 3.24, the GOF and SHO results for the combined data from all experiments by Regier and Carlson (2001) are plotted. Table 3.8 shows the model parameters of the best fit to this whole data set for each model. The $rAVS_{c-o-m}$ model clearly obtains the worst results. The $rAVS_{comb}$ model gets a slightly worse result than the AVS, the $rAVS_{prox}$ and the $rAVS_{w-comb}$ models. For comparison, I have also computed the results for the three competitor models of the AVS model from Regier and Carlson (2001): the BB, PC, and the PC-BB models (see Section 3.1.1). All these models get worse results than most of the (r)AVS models (see Figure 3.24).

To compare the results on all data with the experiment-wise results, I have ranked each model: A model gets the rank 1, if it obtains the best SHO result, the rank 2 if it obtains the second-best SHO result, and so on. If two models are not distinguishable in terms of their SHO result, they get the same rank. For example, the ranks for the results on all data (Figure 3.24; excluding the PC(-BB) models) would be: rank 1: AVS, $rAVS_{prox}$, $rAVS_{w-comb}$; rank 2: $rAVS_{comb}$; rank 3: $rAVS_{c-o-m}$.

Computing the mean rank of each model across all experiment-wise results gives the following ranking: $rAVS_{w-comb}$ (1.26) > $rAVS_{comb}$ (1.63) > $rAVS_{c-o-m}$ (1.84) > AVS (1.89) > $rAVS_{prox}$ (2). Although my ad-hoc

Table 3.8: Model parameter values, nRMSE and correlation of the best fit to all data from Regier and Carlson (2001). λ values for rAVS models are presented in parentheses because they do not change the model outcome (see page 34).

	AVS	rAVS _{prox}	rAVS _{c-o-m}	rAVS _{comb}	rAVS _{w-comb}
λ	0.189	(2.295)	(3.011)	(1.948)	(1.221)
slope	-0.005	-0.003	-0.004	-0.005	-0.004
intercept	0.973	0.946	0.962	0.999	0.943
highgain	0.083	0.076	6.156	0.089	7.497
α	–	–	–	–	0.322
nRMSE	0.073	0.077	0.093	0.078	0.078
R ²	0.972	0.965	0.951	0.966	0.970

ranking method clearly has some caveats (e.g., the number of data points is not considered at all), it provides further support for the rAVS_{w-comb} model. It is interesting to note that despite being ranked as the worst model, the rAVS_{prox} model still obtains a good fit to the whole data set (Figure 3.24). This shows that it is important to use multiple methods to assess model performance.

The experiment-wise model assessments suggest that none of the rAVS variations (except for the rAVS_{w-comb} model) was able to qualitatively accommodate all empirical effects (rAVS_{prox} could not accommodate data from exps. 3, 4, 7; rAVS_{c-o-m} & rAVS_{comb} could not accommodate data from exp. 7). This leaves only two models that are not distinguishable in terms of performance across all data: The AVS model and the rAVS_{w-comb} model. I contrasted the AVS model and the rAVS_{w-comb} model further by generating model predictions on novel stimuli and conducting an empirical study with these stimuli (see Chapter 4).

3.3 DISCUSSION OF EVALUATION OF RAVS VARIATIONS

By comparing the performance of the successful rAVS_{w-comb} model with the performance of the AVS model, I showed that a fundamental (although implicit) assumption of the AVS model seems not to be necessary to model spatial language verification: The direction of the attentional shift from the RO to the LO. The rAVS models challenged this assumption and indeed, the rAVS_{w-comb} model replicates all empirical effects found by Regier and Carlson (2001). Accordingly, the reported results neither favor an attentional shift from the RO to the LO – as assumed by the AVS model and its theoretical background (Logan, 1995; Logan & Sadler, 1996; Logan & Zbrodoff, 1999; Regier & Carlson, 2001) – nor a shift in the opposite direction – as assumed by the rAVS_{w-comb} model motivated by theoretical and empirical research

Based on the existing empirical rating data, neither directionality of the attentional shift can be favored over the other.

(e.g., Burigo & Knoeferle, 2015; Chambers et al., 2002; Franconeri et al., 2012; Roth & Franconeri, 2012). Rather, both directionalities of attention are equally well supported by the simulation results.

The reported model simulations underline the importance of both center-of-mass orientation and proximal orientation for the acceptability of a spatial utterance while they add another factor into the equation: relative distance. The parameter that controls the relative importance of either orientation via relative distance in the $rAVS_{w-comb}$ model, α , was robustly estimated to be around 0.3 when the $rAVS_{w-comb}$ model was fitted to all data from Regier and Carlson (2001). This means that the proximal orientation might become irrelevant for LOs placed with a relative distance greater than 3. For closer LOs, the proximal orientation gets more important the closer the LO is. In the results of the empirical study presented in Section 4.2.1, we will see whether the proposed role of relative distance holds true.

In the preceding chapter, I presented several variations of the rAVS model. All these models are based on the AVS model but reversed the directionality of the shift of attention: Instead of implementing a shift of attention from the RO to the LO like the AVS model, the rAVS models implement an attentional shift from the LO to the RO. Assessing all models on the rating data from Logan and Sadler (1996), Hayward and Tarr (1995), and Regier and Carlson (2001) revealed that one rAVS variation – the rAVS_{w-comb} model – accounts for the existing empirical data as well as the AVS model (measured via GOF, SHO, and replications of qualitative patterns). Given that the rAVS_{w-comb} model and the AVS model implement contrasting directionalities of the attentional shift, the model simulations on the existing empirical data do not favor one implemented directionality of attention over the other.

How could one further contrast the two implemented directionalities in order to decide whether any of the directionalities better describes human processing? My idea was to empirically test predictions that the two models (AVS and rAVS_{w-comb}) make. These predictions stem from the different model mechanisms, which, in turn, are implications of implementing the directionality of the attentional shift (from the RO to the LO or vice versa). Hence, empirically testing model predictions is a test whether any directionality of attention better describes human data than the other. This chapter reports on the generation of the model predictions and the empirical study that tested the generated predictions.

More specifically, based on the different model mechanisms, I have designed two types of stimuli for which the models predict somewhat different outcomes. The stimuli test (i) rAVS_{w-comb}'s mechanism of relative distance between LO and RO and (ii) the influence of asymmetrical ROs. Section 4.1 introduces the two types of stimuli in detail. In Section 4.2, I present the results of an empirical study asking whether humans follow the model predictions for these stimuli. If so, the model making the correct prediction would be supported by the empirical outcome. Moreover, the two models might perform differently on the newly collected empirical data, i.e., one model could be favored over

Two test cases (relative distance & asymmetrical ROs) to compare model predictions implied by implementing contrasting directionalities of the attentional shift.

* Parts of the work presented in Chapter 4 were published in Kluth, Burigo, and Knoeferle (2016a, stimuli, PSP), Kluth, Burigo, and Knoeferle (2016b, PSP, empirical study), Kluth, Burigo, Schultheis, and Knoeferle (2016b, asymmetrical ROs data), Kluth, Burigo, Schultheis, and Knoeferle (2017, relative distance data), and Kluth, Burigo, Schultheis, and Knoeferle (2019, PSP, empirical study). This text extends on the already published details and presents a comprehensive overview of all analyses.

the other. This is why Chapter 5 presents a thorough analysis of the model performance on the stimuli and the collected data.

4.1 PREDICTIONS

4.1.1 *Relative Distance*

The $rAVS_{w-comb}$ model explicitly uses the relative distance between LO and RO in its mechanism of weighting the proximal orientation and the center-of-mass orientation: The closer an LO is to an RO, the more important the proximal orientation gets. On the other hand, the rating for a more distant LO is more strongly influenced by the center-of-mass orientation. In particular, $rAVS_{w-comb}$'s weighting mechanism is sensitive to *relative* and not *absolute* distance, where relative distance is (roughly) defined as absolute distance divided by the size of the RO (formulated more precisely in Equation 3.5 on page 37). Accordingly, the $rAVS_{w-comb}$ model predicts different ratings for two spatial configuration where the LOs have the same *absolute* distance to the RO but different *relative* distances. Such configurations can be seen in Figure 4.1. For both LOs in Figure 4.1a and Figure 4.1b, the absolute distance to the rectangular RO is equal. However, due to the different sizes of the ROs, the relative distance of the LO to the RO is smaller in Figure 4.1b ($1.5/3.0 = 0.5$) than it is in Figure 4.1a ($1.5/1.5 = 1.0$). With lower relative distance, the proximal orientation gets more important. Thus, $rAVS_{w-comb}$'s vector points more towards the proximal point on top of the RO which leads to a lower angular deviation and in turn to a higher rating. Accordingly, the $rAVS_{w-comb}$ model computes a higher rating for the LO with the lower relative distance (Figure 4.1b) compared to the LO with the larger relative distance (Figure 4.1a).

rAVS_{w-comb}: the lower the relative distance, the higher the rating.

The AVS model, on the other hand, does not explicitly state an effect of relative distance. The AVS model has four ways to account for distance effects: The height component, the λ parameter, the σ variable and the vector sum. Since the height component is shared with the $rAVS_{w-comb}$ model, no conflicting predictions emerge from it. The parameter λ controls the attentional width and is freely adjustable, but once λ is fixed it is valid for all LOs. Based on a fixed λ , the two attentional distributions for the two configurations in Figure 4.1 are exactly the same. The variable σ stands for the absolute distance between the LO and the focal point and also controls the attentional width: LOs farther away result in a greater attentional width than LOs close to the RO. However, because σ only codes *absolute* distance, it is the same for the two configurations in Figure 4.1 (namely 1.5). Taken together, because the attentional distribution that is controlled by λ and σ is the same for both configurations, neither λ nor σ predict different ratings for the two configurations in Figure 4.1.

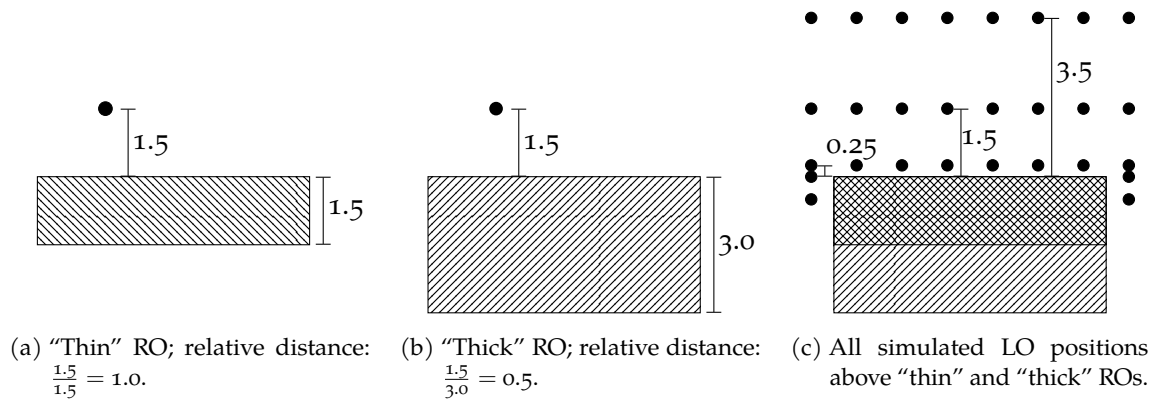
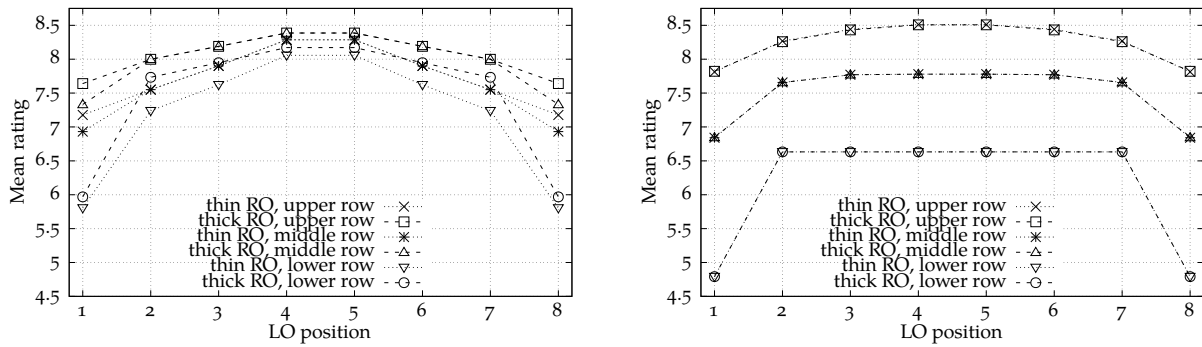


Figure 4.1: Spatial configurations to test the effect of relative distance of the LO to the RO. ROs are filled with different patterns for visualization purposes only.

The last possible source of different ratings is the vector sum: Due to the different number of points in the two ROs, the vector sum in Figure 4.1a consists of less vectors than the vector sum in Figure 4.1b. In particular, the same vectors for the RO in Figure 4.1a also exist for the RO in Figure 4.1b. In Figure 4.1b the number of vectors is twice the number of vectors for Figure 4.1a (because the RO in Figure 4.1b is twice the size than the RO in Figure 4.1a). However, the additional vectors have a lower attentional weight. Due to this difference in the number of vectors, the final direction vector that is compared to canonical upright is different for Figure 4.1a than for Figure 4.1b. Thus, the vector sum predicts different ratings for the two ROs. However, to what extent the ratings should differ is not immediately clear. If the attentional width is large, the additional vectors in Figure 4.1b get more attentional weight and thus might change the final direction vector considerably. If, on the other hand, the attentional width is small, the final direction vector changes only marginally. The AVS model accordingly predicts higher differences in ratings for distant LOs than for close LOs, because LOs that are far away result in a greater attentional width (due to a greater σ).

Since the computation of the vector sum is hard to grasp, it is also difficult to state a clear prediction from the AVS model. Based on the above reasoning, I would expect a subtle difference in the two ratings. But is this subtle difference distinguishable from the difference that the $rAVS_{w-comb}$ model predicts? Are the configurations shown in Figures 4.1a and 4.1b suitable for contrasting the two models? To explore this, I have generated ratings with each model for the three rows of LOs shown in Figure 4.1c. I have used a fixed set of parameters: The parameters of the best fit to all data from Regier and Carlson (2001) for the respective model (see Table 3.8). The ratings are plotted in Figure 4.2. The AVS model computes the same patterns of rating regardless of the

It is unclear whether the AVS model predicts an effect of relative distance.



(a) rAVS_{w-comb} ratings for LOs above thin and thick rectangle (see Figure 4.1). (b) AVS ratings for LOs above thin and thick rectangle (see Figure 4.1).

Figure 4.2: Qualitative comparison of (a) rAVS_{w-comb}-generated ratings and (b) AVS-generated ratings for LOs above thin and thick rectangle (see Figure 4.1c). For data generation, I have used model parameters from best fit to all data from Regier and Carlson (2001, Table 3.8).

RO, i.e., regardless of the relative distance of the LO (see Figure 4.2b).¹ It seems that only the absolute distance matters for the AVS model – at least for the model parameters used here. On the other hand and as expected, the rAVS_{w-comb} model predicts higher ratings for the LOs above the thick RO compared to LOs above the thin RO, as can be seen in Figure 4.2a: ratings for same rows are higher for the thick RO than for the thin RO. Different than the AVS model, the rAVS_{w-comb} model predicts almost the same ratings for the upper two rows above the same RO – at least with the parameters used to generate these ratings.

To investigate the whole space of predictions from each model, I have applied the ‘Parameter Space Partitioning’ method (PSP, Kim, Navarro, Pitt, & Myung, 2004; Pitt et al., 2006). This method generates qualitative model predictions for each set of parameters and thus quantifies the numbers of different model predictions across the whole parameter space. Before I present the PSP results in Section 4.1.3, I introduce a second set of stimuli that potentially elicits different model predictions, too.

4.1.2 Asymmetrical ROs

The second source of model predictions are asymmetrical ROs with a flat top. While the AVS model in principle is able to reflect arbitrary RO shapes within its vector sum, the rAVS_{w-comb} model simplifies the geometry of the RO using two points only: one point on top of the RO and the center-of-mass of the RO. While the center-of-mass incorporates

¹ There are tiny rating differences across different ROs for corresponding LO placements (i.e., LOs in the same row at the same position). However, these differences are smaller than 0.01, and thus not visible in the plot.

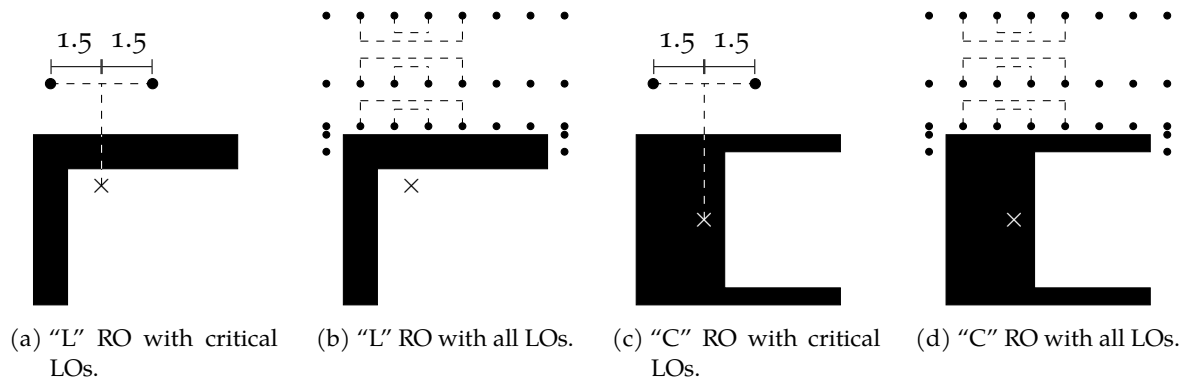


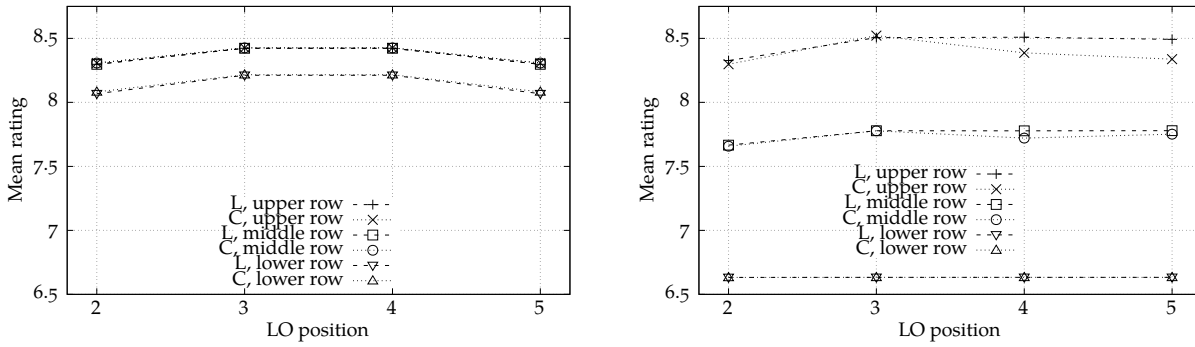
Figure 4.3: Spatial configurations to test the effect of asymmetrical ROs. (a, b): "L" RO, (c, d): "C" RO. (a, c): two critical LO positions, (b, d): all simulated LO positions with critical LO-pairs connected with a dashed line. \times : center-of-mass of RO.

asymmetries of the object, it does so in a more condensed way than the vector sum. I have developed two asymmetrical ROs with corresponding LO positions, shown in Figures 4.3a and 4.3c, for which the $rAVS_{w-comb}$ model predicts equal ratings. The symbol \times in Figure 4.3 depicts the center-of-mass of each RO. As visualized with dashed lines, the critical LOs in Figures 4.3a and 4.3c have the same horizontal and vertical distance from the center-of-mass – that is, both LOs have the same center-of-mass orientation. Nevertheless, both left LOs are directly above the part of the ROs that has more mass. In contrast the right LOs are located above the "cavity" of the ROs. The $rAVS_{w-comb}$ model cannot represent the asymmetry of the ROs (except that it is already integrated in the center-of-mass which is shifted to the left compared to symmetrical ROs). Accordingly, the $rAVS_{w-comb}$ model predicts exactly the same rating for the two LOs in Figures 4.3a and 4.3c – despite the asymmetrical distribution of mass directly below the LOs.

The AVS model, on the other hand, seems to predict different ratings for these LOs. The AVS model represents the whole RO with its vectors. In particular, the vertical bar of the RO in Figure 4.3a is represented in detail. This should lead to higher rating for the left LO in Figure 4.3a compared to the right LO. This is because for the left LO, more vectors on the vertical bar are closer to canonical upright – compared to the right LO, where the same vectors on the RO's vertical bar have a greater deviation from upright vertical. This effect is the same for the RO in Figure 4.3c.

Furthermore, the attentional distribution changes for the two LOs. While the attentional width stays the same for both LOs (σ is equal across configurations with equal LO-RO distance), the attentional focus is not the same. So, for the left LO in Figure 4.3a, the vertical bar of the RO gets more attention and hence should have a greater impact on the direction of the final vector and thus on the rating. For the computation

The $rAVS_{w-comb}$ model predicts no difference in ratings for LOs above asymmetrical ROs, placed with equal distance to the center-of-mass.



(a) rAVS_{w-comb} ratings for critical LOs above L and C RO (see Figures 4.3b and 4.3d). (b) AVS ratings for critical LOs above L and C RO (see Figures 4.3b and 4.3d).

Figure 4.4: Qualitative comparison of (a) rAVS_{w-comb}-generated ratings and (b) AVS-generated ratings for critical LOs above L and C RO (see Figures 4.3b and 4.3d). For data generation, I have used model parameters from best fit to all data from Regier and Carlson (2001, Table 3.8).

of the rating for the right LO, however, the vertical bar might receive very little attention (depending on the attentional width) and thus might play almost no role for the rating.

Again it is hard to come up with a prediction for the AVS model because of the flexibility of the attentional vector sum. However, it seems fair to say that the AVS model at least intuitively predicts different ratings for the two LOs in Figure 4.3a and 4.3c (based on the asymmetrical distribution of mass directly below the LOs). In contrast, the rAVS_{w-comb} model predicts no differences by definition.

This also emerges in the qualitative comparison of both models on these two ROs. Figure 4.4 shows the ratings for six critical pairs of LOs above the ROs (see Figure 4.3b and 4.3d). For the plotted ratings, I again used the parameters from the best fit to all data from Regier and Carlson (2001), see Table 3.8. The rAVS_{w-comb} model computes the exact same rating for LOs in the same row at position 2 and 5 or at position 3 and 4, respectively (see Figure 4.4a). Moreover, all positions in the upper and the middle row get virtually the same rating – the ROs do not matter (at least for this parameter setting). This is also true for the lower row, which gets an overall lower rating compared to the other two rows, but almost the same ratings compared across ROs. The AVS model computes a more complicated rating pattern (see Figure 4.4b). First of all, the ratings in the upper two rows at positions 3 and 4 for the C-shaped RO are different (but only to a very small amount). Next, the ratings in the upper two rows for positions 2 and 5 are different for the L shaped RO (but again it is a tiny difference). The ratings for the lower row are all equal.

The AVS model seems to predict different ratings for LOs above asymmetrical ROs, placed with equal distance to the center-of-mass.

This comparison has shown that the AVS model is able to compute distinct ratings with an input for which the $rAVS_{w-comb}$ computes identical ratings (due to its definition). Whether the ratings computed by the AVS model stay different for other sets of parameters, however, remains unclear. It might well be that with a different parameter set the AVS model predicts no difference at all – or a bigger difference than that of Figure 4.4b. To properly investigate the full range of model predictions, I have applied the PSP method for these stimuli, too. The next section introduces this method and discusses its results on both sets of stimuli (asymmetrical ROs and differently sized rectangles).

4.1.3 *Parameter Space Partitioning*

METHOD Kim et al. (2004) and Pitt et al. (2006) introduced a method called ‘Parameter Space Partitioning’ (PSP). The PSP method discovers all *qualitatively* different output patterns a model can generate by evaluating the model outcome throughout its parameter space. However, due to the size of the parameter space, simulating the model with all possible parameter sets is a time-consuming process. In fact, it is often not feasible (depending on models, stimuli, hard-, and software). This is why Kim et al. (2004) and Pitt et al. (2006) developed the PSP algorithm as a tool to explore the space of the free model parameters in a matter of minutes or hours – in contrast to days or weeks that a naïve complete enumeration of the parameter space would take. Internally, the PSP algorithm is a Markov Chain Monte Carlo algorithm that searches through the parameter space of the models. I have constrained the model parameters for the PSP (i.e., the boundaries of the parameter space) in the same way as for the parameter estimation used for GOF and SHO (see Equations 3.15–3.19 on page 42). I have used the MATLAB implementation that Pitt et al. (2006) made available² with GNU octave (an open source MATLAB clone, Eaton, Bateman, Hauberg, & Wehbring, 2015).

INPUT FOR THE PSP METHOD To use the PSP algorithm one must first define what a qualitative pattern looks like. In my case, a qualitative pattern describes the relationship of mean ratings for at least two different spatial configurations. Given two different LO-RO pairs, there are three possible qualitative rating patterns: (i) The first configuration is rated higher than the second configuration (coded as “+”), (ii) the first configuration is rated lower than the second configuration (coded as “-”) or (iii) both configurations get the same rating (coded as “0”). But when should two mean ratings be considered to be equal? In empirical rating studies, statistical analyses are used to investigate whether em-

“+”: *first* > *second*
 “-”: *first* < *second*
 “0”: *first* = *second*

² <http://faculty.psy.ohio-state.edu/myung/personal/psp.html>. I slightly changed the source code due to compatibility issues. The changed source code is available under Kluth (2018).

irical ratings for two different configurations are significantly different. Since the cognitive models only compute mean ratings, we cannot use such techniques (but see Section 5.7 for a model extension that allows to generate individual data).

Previous spatial language acceptability rating studies provide valuable information regarding the magnitude of difference in mean ratings necessary for statistically significant effects. Those studies investigated different aspects of spatial language with varying effect sizes. This is why I considered the studies' significant differences only as approximate benchmarks and distilled three different "equality of ratings" thresholds (t_e) for the PSP algorithm. All following differences are differences in mean ratings (but see Liddell & Kruschke, 2018, why it is problematic to interpret ordinal data as metric). Regier and Carlson (2001) found that "a critical difference of [0.17, 0.2, 0.3, 0.7 was required] for significance" (exps. 1 and 2, p. 282; exp. 4, p. 285; exp. 6, p. 288). They used a rating scale from 0–9. Carlson-Radvansky et al. (1999) used a scale from 1–7 and in their experiment a difference of 0.3 was significant. Hörberg (2008, p. 208) also used a rating scale from 1–7 and his experiment required a difference of 0.57 for significance. Burigo et al. (2016) used a rating scale from 1–9 and found the following differences to be significant in their second experiment (p. 11): 0.32 and 0.39. Following these benchmarks, I have used the following three thresholds t_e for equality of ratings in the PSP algorithm: $t_e \in [0.1, 0.5, 1.0]$.

In my PSP analysis, I included the following three comparisons (see Figure 4.5): two LOs above the asymmetrical C (with equal center-of-mass orientation), two LOs above the asymmetrical L (with equal center-of-mass orientation), one LO above the thin rectangle versus one LO above the tall rectangle. Before presenting the PSP results, let us revisit the "intuitive" model predictions for these stimuli (see Sections 4.1.1 and 4.1.2). For the asymmetrical ROs, the $rAVS_{w-comb}$ model clearly predicts no difference for two LOs placed at the same horizontal and vertical distance from the center-of-mass. Using the coding explained above, the $rAVS_{w-comb}$ model predicts a "0" for the first two comparisons. The AVS model, however, seems to predict a higher rating for the left LOs above the asymmetrical ROs (compared to the right LOs), i.e., higher ratings for the LOs that have more mass of the ROs directly below it. Accordingly, the AVS models predicts a "+" for the first two comparisons.

For the relative distance test case, the $rAVS_{w-comb}$ model predicts lower ratings for ROs with less height (larger relative distance) compared to ROs with greater height (lower relative distance). That is, for the last comparison, the $rAVS_{w-comb}$ model predicts a "-". The prediction of the AVS model for the relative distance case is unclear. Taken together, the $rAVS_{w-comb}$ model predicts the three-digit pattern "00-" (no difference for LOs above the asymmetrical ROs and a lower rating for the LO above the thin rectangle compared to the LO above the tall

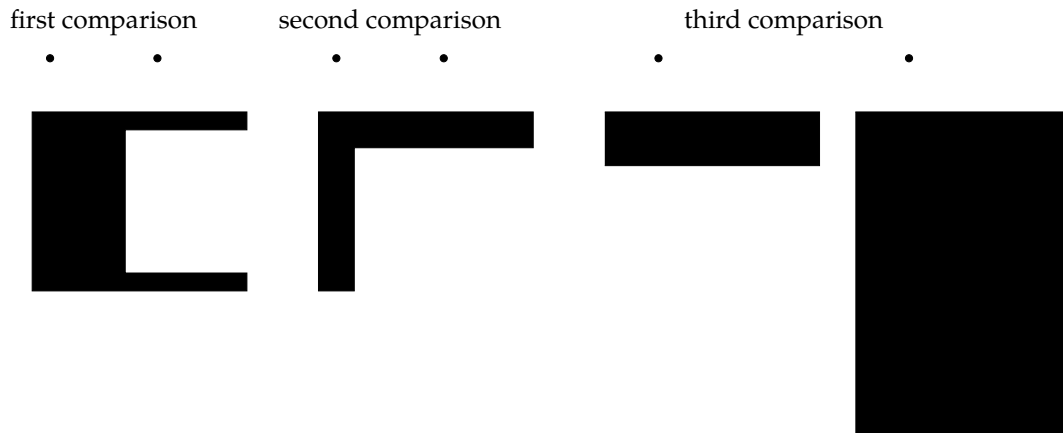


Figure 4.5: ROs and LOs used as input for the PSP method. First comparison was between the two LOs above the C RO, second comparison between the two LOs above the L RO, third comparison between the two LOs above the thin vs. the tall rectangle.

rectangle), whereas the AVS model predicts the pattern “++?” (higher rating for LOs above the part of the asymmetrical ROs that has more mass and an unclear prediction for the relative distance condition).

RESULTS The main outcome of one PSP run are estimates of volumes in the parameter space for each qualitative pattern a model generates. I ran the PSP algorithm three times for every model and equality threshold t_e . The mean estimates of relative³ volumes from the three runs are plotted in Figure 4.6, separately for each threshold t_e .

Throughout all thresholds t_e , the $rAVS_{w-comb}$ model only generates 2 out of 27 possible rating patterns: “000” and “00-”. The latter pattern was “intuitively” predicted beforehand and also occupies the majority of the parameter space for all equality thresholds. The reason that the volume of pattern “000” increases with increasing threshold t_e is the following: If two mean ratings differ by, say, 0.3 they are considered as not equal with $t_e = 0.1$ but as equal with $t_e = 0.5$.

The AVS model generates 3 out of 27 patterns for $t_e = 1.0$ (Figure 4.6c): The same two patterns that the $rAVS_{w-comb}$ model generates (“000” and “00-”, with different proportions than the $rAVS_{w-comb}$ model) and the additional pattern “0-0” with a small amount of estimated volume. The smaller the threshold t_e , the more patterns are generated by the AVS model: For $t_e = 0.1$ and $t_e = 0.5$ (Figures 4.6a and 4.6b), the AVS model generates 7 out of 27 possible patterns. For $t_e = 0.5$ (Figure 4.6b), 4 of these patterns together occupy less than 15% of the parameter space. The rest of the parameter space is occupied by the same two patterns that the $rAVS_{w-comb}$ model generates (“000” and

The PSP analysis confirms the “intuitive” predictions for the $rAVS_{w-comb}$ model.

³ The estimated parameter space volumes returned by the PSP implementation do not sum to 100%. I extrapolated the estimated volumes to cover the full parameter space.

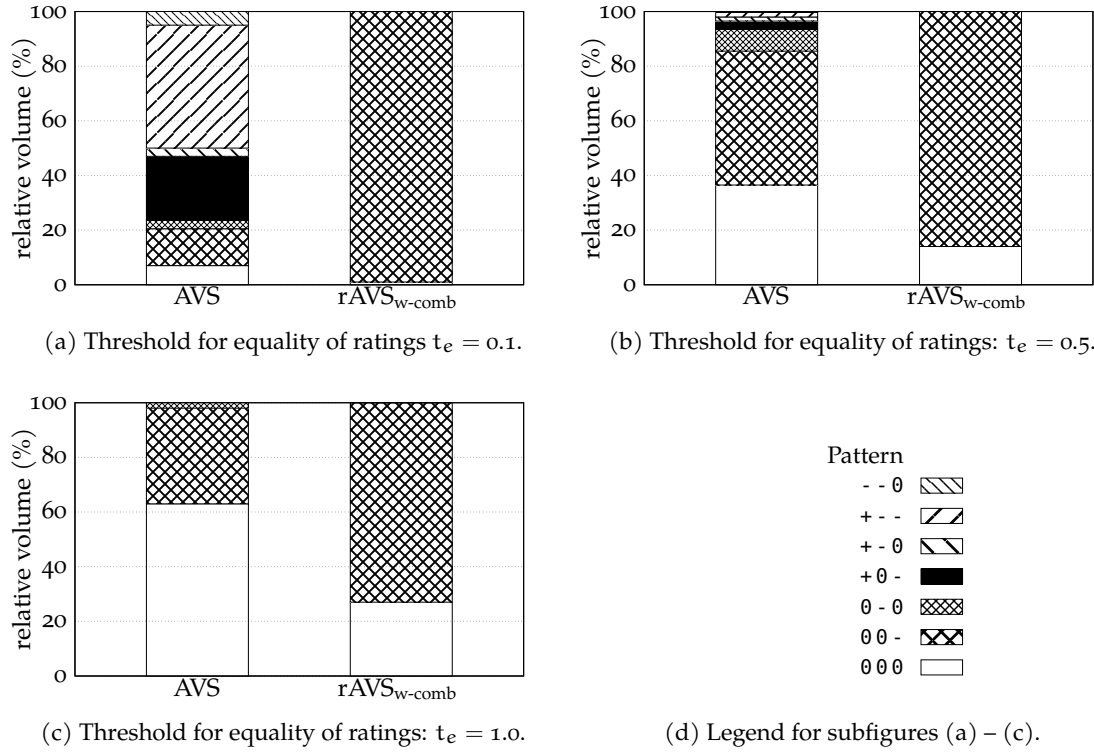


Figure 4.6: PSP results: Estimations of relative volumes in parameter spaces of the models covered by distinct qualitative patterns for spatial configurations depicted in Figure 4.5. Subfigure (d) shows legend for all plots (a)–(c). First symbol in pattern: rating difference for LOs above C RO; second symbol: rating difference for LOs above L RO; third symbol: rating difference for LOs above thin vs. tall rectangle. Two mean ratings were considered to be different if they differed by more than (a) $t_e = 0.1$ (b) $t_e = 0.5$ or (c) $t_e = 1.0$. Mean estimates of three PSP runs are plotted.

The PSP analysis does not confirm the “intuitive” predictions for the AVS model.

“00-”). For $t_e = 0.1$ (Figure 4.6a), however, these 2 patterns occupy only approximately 20% of the parameter space, whereas the 2 patterns “+0-” and “+ - -” are generated throughout almost 70% of the parameter space. Thus, changing the value of t_e obviously also changes the qualitative predictions of the AVS model – despite the fact that all t_e are reasonable in terms of previous research. More interestingly, the “intuitive” prediction stated above (“++?”) does not show up at all in the PSP results for the AVS model. This suggests that it is difficult to intuitively reason with the AVS model.

Taken together, the AVS and rAVS_{w-comb} models generate the same qualitative patterns for the two test cases (asymmetrical ROs, relative distance). However, while the rAVS_{w-comb} model generates only one additional, closely related qualitative pattern, the AVS model generates up to six more distinct qualitative patterns. This makes the AVS model more flexible than the rAVS_{w-comb} model (for a more thorough discussion of model flexibility see Section 5.4). Although the models share

some of their predictions, it is worth to gather empirical data in order to investigate whether humans follow these predictions. Additionally, on the newly collected data the two models could still perform differently in terms of quantitative fits (PSP only computes qualitative patterns). This could help to distinguish between both models. To this end, the next section presents an empirical study with the same stimuli as used in the PSP analysis.

4.2 EMPIRICAL STUDY

The AVS model and the $rAVS_{w-comb}$ model are not distinguishable on the existing empirical data (Hayward & Tarr, 1995; Logan & Sadler, 1996; Regier & Carlson, 2001, see Chapter 3). However, the two models are predicting somewhat different outcomes for particular displays (see Section 4.1). This is why I conducted an empirical study in order to test these predictions and to collect data that might help in distinguishing the two models.⁴ The main goal of the study was to provide data on which the AVS and the $rAVS_{w-comb}$ model would perform differently. Besides specifically testing the model predictions for the relative distance and the asymmetrical ROs test cases, I also analyzed whether the experiment replicated known effects from the literature (effect of superior vs. inferior preposition on rating and reaction time; effect of grazing line on rating). The remainder of this chapter introduces the empirical study (materials, procedure, and statistical method: the next pages) and the analyses of the data (ratings, eye movements, reaction times: Sections 4.2.1–4.2.3). The following Chapter 5 applies a range of model evaluation techniques on the data and stimuli from the study with the goal of distinguishing the AVS model from the $rAVS_{w-comb}$ model.

I designed the study as a rating study similar to Logan and Sadler (1996) and Regier and Carlson (2001): First a spatial sentence is shown (like “The dot is above the object”) and afterwards a spatial configuration is shown for which the sentence acceptability should be rated. Although this design does not allow to investigate visual attention during the processing of the unfolding spatial language utterance, eye tracking provides valuable data about the deployment of visual attention in spatial language rating tasks. This is why I tracked participants’ eye movements during the inspection of the spatial configuration.

MATERIALS In the empirical study, I have used the same spatial configurations as for the PSP analysis and included some additional stimuli. I added the mC and mL ROs as vertically flipped versions of the C and L ROs to obtain a left-right balance. Furthermore, I added two more rectangular ROs to test the influence of relative distance. Figure 4.7 depicts all ROs used in the study and also presents their code names.

⁴ Michele Burigo was a great help in setting up and conducting the reported experiment.

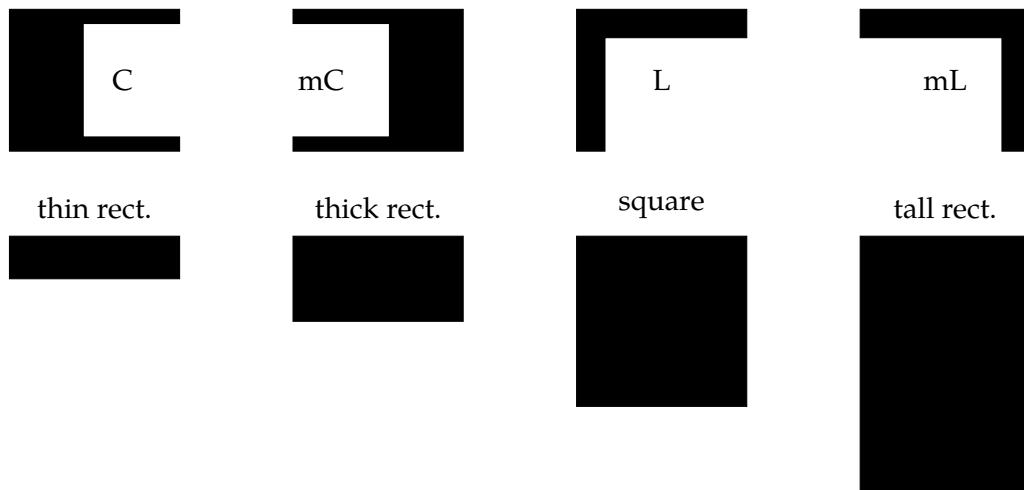


Figure 4.7: All ROs and their code names used in the empirical rating study.

For two example ROs, Figure 4.8 shows all LO placements and the respective row and column coding. I placed 28 LOs above each RO (rows R1–R5) and 28 LOs below each RO (rows R6–R10). Out of these 2×28 LOs, I placed 2×4 LOs on or slightly below/above the grazing line (rows R4–R7). For the 28 LOs above each RO (rows R1–R5), participants had to rate the acceptability of the German description “Der Punkt ist über dem Objekt” (“The dot is above the object”). For the 28 LOs below each RO (rows R6–R10), participants read the sentence “Der Punkt ist unter dem Objekt” (“The dot is below the object”). I horizontally flipped the asymmetrical L and mL ROs for the 28 LOs below these ROs such that the RO was always facing the LO with a flat surface on the bottom/top (see Figure 4.8).

Participants saw each RO-LO combination exactly once, i.e., only one RO and only one LO were present at the same time. No additional information was provided in the spatial configuration displays (no row or column numbers, no center-of-mass, etc.). I placed each RO such that its center-of-mass coincided with the center of the screen. The LOs were then placed accordingly (relative to the borders of the RO). Taken together, this rating study consisted of $8 \text{ ROs} \times 28 \text{ LOs} \times 2 \text{ prepositions} = 448$ items. Due to the length of the experiment, no fillers were added.

Participants sat in front of a computer monitor (22 inches, 1680×1050 pixel) with a distance of 80 cm. They had to use a chin rest. During the display of the spatial configurations, participants’ right eye was tracked with a desktop mounted eye tracking system (EyeLink 1000, SR Research). I used the software “Experiment Builder” (version 1.10.125, SR Research) to program the experiment. All files to recreate the experiment, the raw result files, and source code with analyses are published under Kluth (2018). This study was approved by the ethics committee of the University of Bielefeld under the number 2015-126.

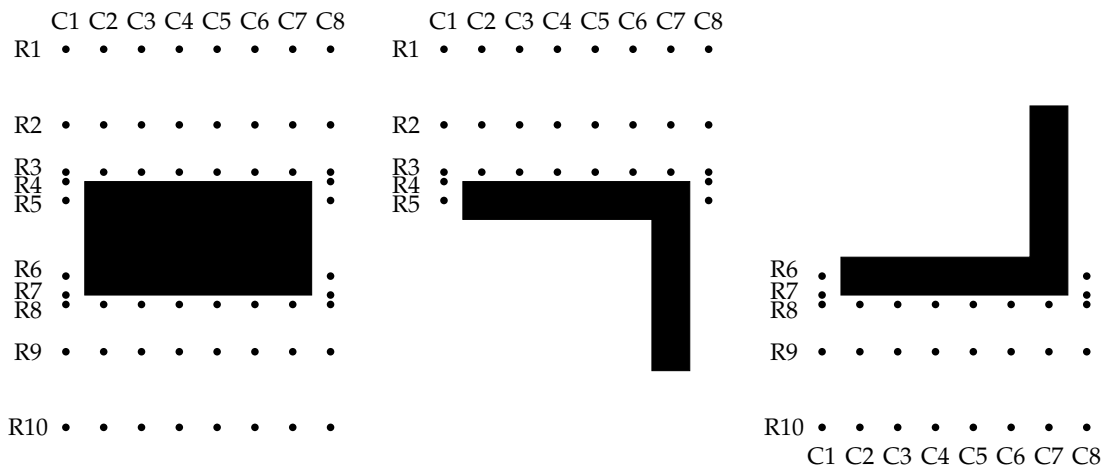


Figure 4.8: LO placements with row and column coding for two example ROs. Rows R1–R5 were presented with *über* (above), rows R6–R10 were presented with *unter* (below).

PROCEDURE I recruited 34 participants (19 females), aged from 18–34 (mean age: 23.79). Most of the participants were either students at the University of Bielefeld or the University of Applied Sciences Bielefeld. They were paid 6 € for participation. The study took approximately 45 minutes. After participants completed a general questionnaire and made themselves comfortable in front of the computer monitor, they read an introductory text (see Appendix B). In the text, they were told that they would see sentence-picture pairs and that they had to rate the picture according to how well it is described by the preceding sentence. To do so, they should use the number keys 1–9 above the letters on a standard keyboard. Here, 1 means “The sentence does not describe the picture at all” and 9 means “The sentence describes the picture very well”. The text encouraged participants to use the whole rating range. After the eye tracker was calibrated, participants rated four practice trials (with different, non-critical ROs). Thereafter, participants rated all 448 items in pseudo-random order (with the possibility to make breaks in between). Pseudo-randomization was done with the only constraint that the same RO should not appear twice in a row. To rate a sentence-picture pair (see also Figure 4.9), participants read the sentence “Der Punkt ist über/unter dem Objekt” (“The dot is above/below the object”, only one preposition per sentence) and pressed space after reading it. Thereafter, one RO and one LO appeared on the screen and were visible until participants responded with their rating. Reaction time was measured from the onset of the spatial configuration until the key press of the rating. Eyes were tracked during the display of the spatial configuration.

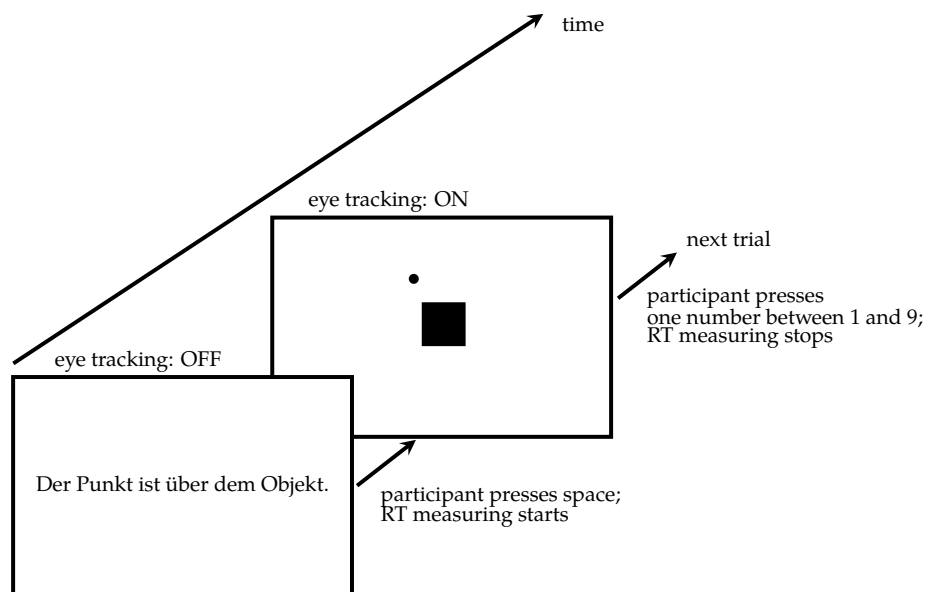


Figure 4.9: Schematic visualization of a single experimental trial. Not to scale.

Method of Data Analysis

The NHST framework is severely flawed.

Bayesian data analysis provides practical and theoretical benefits over NHST.

I conducted all following data analyses using the Bayesian framework. I did this because there is growing consensus that the classical ‘Null Hypothesis Significance Testing’ (NHST) framework focusing on the significance of an effect given a corresponding $p < 0.05$ is severely flawed (e.g., Dienes, 2011; Gigerenzer, 2004; Kruschke, 2013; Lindley, 1993; Wagenmakers, 2007; Wagenmakers et al., 2018). The proposal to use confidence intervals instead of p values (so called “new statistics” by Cumming, 2014) shows that psychologists have become aware of the problems of p values. Confidence intervals, however, are still operating within the same NHST framework as p values and accordingly suffer from similar problems (e.g., Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Kruschke, 2013). Kruschke and Liddell (2018) claim that the goals of the “new statistics” can better be reached by using Bayesian methods. Furthermore, NHST can only be used to answer questions like “How probable are the data given theory T?”. Bayesian data analysis, in addition, allows to answer questions like “How probable is theory T given the data?” – presumably, this is what most researchers are more interested in (Dienes, 2011).

This is why I used the Bayesian framework for my data analyses. It provides a coherent framework that can be used for hypothesis testing (including eventually accepting the null hypothesis), as well as parameter estimation (in order to know the size of the effect). More-

over, I can integrate prior knowledge (from previous research or model simulations) into the statistical models of the data.⁵

A very brief introduction to Bayesian data analysis is the following: First, we identify the nature of our data. Then, we formulate a statistical model of these data that specifies which probability distribution the data should follow. Moreover, the statistical model tries to predict the data as a function of the experimental conditions. Thus, the model has parameters describing the relationship of predictor variables to predicted variables (e.g., a positive slope parameter means a rise of the outcome variable) and provides a likelihood function (how likely are the data given the model with specific parameter values). Incorporated in this statistical model is our prior knowledge about the effect via prior probability distributions over the parameter space. If there is no such knowledge, the prior probabilities can be specified to be vague.

Using Bayes rule and computing power, we can estimate the posterior probability distribution for the model parameters which consists of the most probable values for the parameters given the data. Since the model and its parameter distributions describe our theory about the data, the posterior distributions of the parameters give an answer to the question “How probable is theory T given the data?”. For an accessible introduction to Bayesian data analysis see the annotated reading list by Etz, Gronau, Dablander, Edelsbrunner, and Baribault (2018). The textbook by Kruschke (2015) provides a comprehensive hands-on tutorial on Bayesian data analysis.

In the Bayesian framework, I used generalized linear multilevel regression models to describe the data, “of which ANOVA, t-tests, linear and logistic regression, χ^2 , and hierarchical loglinear models are examples (with likelihood based inference lying at the heart of all of these)” (Altmann, 2007, p. 4). Accordingly, all following results can be interpreted much like results from one of these more traditional NHST analyses. Since I have designed the study using a repeated measurement design, I included subjects in the group-level of the multilevel regression models to account for inter-subject variability.

To model acceptability ratings (an ordinal outcome variable, i.e., discrete and ordered), I used an ordinal regression model (e.g., Kruschke, 2015, Chapter 23; Liddell & Kruschke, 2018). This type of regression assumes a latent metric distribution and predicts the outcome variable as follows (cf. Figure 5.9c and 5.9d on page 141): On the latent metric distribution, $r - 1$ thresholds are estimated (where r is the number

⁵ However, I could inform my analyses only with rather uncertain prior distributions. This is because the predicted variable (rating) in my study is an ordinal outcome. Accordingly, I applied ordinal regression. In contrast, regression coefficients from the literature are derived from interpreting acceptability ratings as a metric variable (a suboptimal analysis of ordinal data, Liddell & Kruschke, 2018). These regression coefficients are not directly compatible with the coefficients of an ordinal regression. This is why I could only use the qualitative trends from earlier regressions but not the exact quantitative results.

of ratings; in my study $r = 9$) “cutting” the latent distribution into r intervals. The cumulative probability density in each interval of the latent distribution is the probability that the corresponding rating is the outcome. In the statistical models presented here, the latent metric distribution is the logistic distribution since I used the logit link function (Kruschke, 2015, p. 435 ff.). For the analysis of the spatial distribution of fixations, I used (multivariate) Gaussian distributions. To model reaction times, I used the exponentially modified Gaussian distribution, a common choice for reaction time analysis (Dawson, 1988; Van Zandt, 2000).

When comparing two different statistical models fitted to the same data, I applied the leave-one-out cross-validation method (LOO) proposed by Vehtari, Gelman, and Gabry (2017, see also Gelman, Hwang, & Vehtari, 2014). The LOO method measures how well a statistical model fits a data set while it considers the effective number of model parameters to control for over-fitting. Like the SHO method, the LOO method is a cross-validation approach. However, different from the SHO method that was primarily designed to assess cognitive models, the LOO method especially focuses on the evaluation of Bayesian statistical models – including readily available implementations in the statistical software R (R Core Team, 2016, package `loo`, Vehtari, Gelman, & Gabry, 2016).

As a measure of uncertainty of the statistical model, I report 95% credible intervals (CI) of regression parameters. In these CIs, 95% of the mass of the posterior density is located. That is, there is a 95% probability that the posterior value of the regression parameter lies in this interval. In the following analyses (contrasting experimental conditions), I mostly discuss whether the regression parameter ‘slope’ is credibly different from zero (and how large it is). For ordinal regression models (used for the analysis of the rating data), the slope parameter denotes changes of the latent metric distribution with respect to experimental conditions (i.e., the values of the predictor variables). If the slope is credibly different from zero, there is a high probability that the empirical ratings from two experimental conditions are credibly different from each other. The larger the slope, the higher the difference in ratings. However, although the slope seems to be tied to the rating scale (1–9), one cannot directly interpret it on the scale of the ratings. For such interpretations one needs to consider the cumulative probabilities which allow to make statements like “In condition X, ratings 7–9 had a 70% probability whereas in condition Y they only had a 30% probability”. Where appropriate, I discuss the analyses in this way.

In terms of software, I extracted a trial report (containing acceptability ratings) and a fixation report (containing fixations) with the software “Data Viewer” (version 1.11.900, SR Research). With these reports, I used R (R Core Team, 2016) for all further analyses. Specifically, I used the R package `brms` (Bürkner, 2017), a convenient frontend for

the R package `rstan` (Stan Development Team, 2016) that is in turn an interface for Stan. Stan provides computational methods to sample from posterior parameter distributions. All source code for the analyses and plots is available from Kluth (2018).

OVERVIEW OF RESULTS I present the data and corresponding analyses in the following order: First, I analyze the human-derived acceptability ratings, asking whether I could replicate known effects. I move on with detailed analyses of ratings for the two specific test cases that motivated this study: asymmetrical ROs and relative distance. Thereafter, I present two analyses of the eye-movement data: investigating the role of AVS's assumed attentional focus point *F* and exploring gaze patterns for processing the asymmetrical ROs. Finally, I analyze the reaction time data mostly for the sake of completeness.

Rating analyses:
pages 91–104
Eye-movement analyses:
pages 104–110
RT analyses:
pages 110–111

4.2.1 Results: Acceptability Ratings

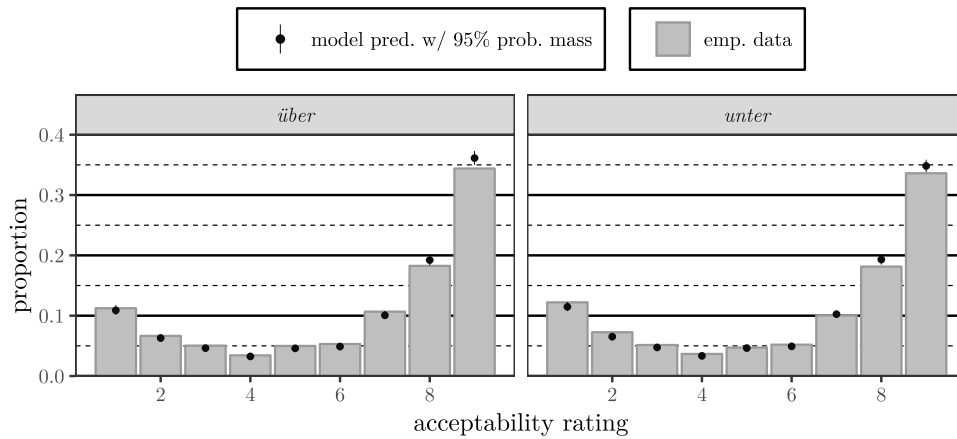
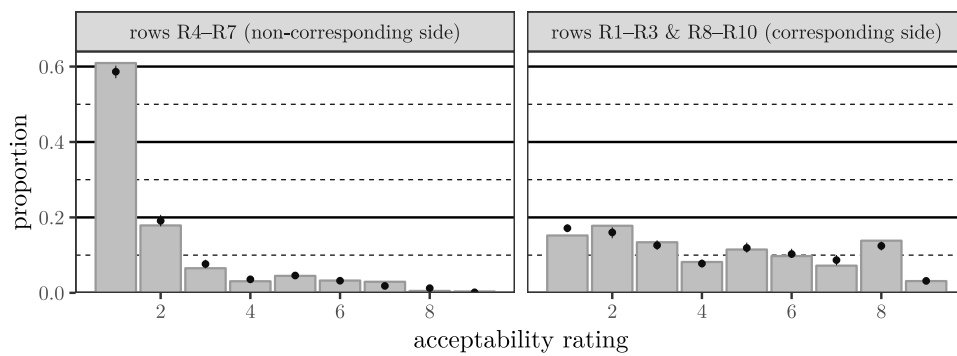
This section reports analyses of the acceptability ratings in three parts. First, I analyzed whether the study replicated known effects. Next, the asymmetrical ROs test case is analyzed. This is followed by an analysis of the relative distance test case.

Acceptability Ratings: Replications

To gain further support for the overall validity of the study, I analyzed whether the collected data replicate effects already established in the literature. Researchers found that people rate superior prepositions (like *über*, *above*) higher than inferior prepositions (like *unter*, *below*; e.g., Burigo & Coventry, 2005; Burigo et al., 2016; Carlson & Logan, 2001; somewhat mirroring the faster response times for superior prepositions compared to inferior prepositions, see page 110). In order to test this finding on the data presented here, I computed an ordinal regression model that predicts rating from preposition. Based on the data from expts. 2 and 3 from Burigo et al. (2016), I estimated the mean of the Gaussian prior distribution of the slope regression parameter as $\mu = -0.11$. Given that Burigo et al. (2016) conducted their study in English and I conducted my study in German, I set the standard deviation of the slope's prior distribution to a relatively large value of $\sigma = 0.2$. This prior distribution favors the effect of lower ratings for *unter* than for *über* while it also puts considerable probability density to a potential null or reversed effect.

Running the regression model with this prior confirmed the effect: People credibly gave lower ratings for *unter* than for *über* ($\beta_{\text{unter}} = -0.0581$, 95% CI $[-0.1152, -0.0006]$). As already evident from the small value of the regression parameter, this is a small effect (see Figure 4.10a): For *über*, participants chose the highest rating 9 with 1% more probability than for *unter*. Recomputing the same regression model with

Higher ratings for über than for unter.

(a) Contrasting *über* (above) and *unter* (below).

(b) Contrasting LOs on both sides of the grazing line.

Figure 4.10: Empirical rating distributions and fits of Bayesian ordinal regression models (computed with 100 samples from the posterior distribution) visualizing the effect of (a) the preposition and (b) the grazing line. Both Bayesian models were instantiated with prior information from earlier research.

brms's default prior (implemented to be non-informative) gives the same probability for choosing rating 9, although the regression parameter is estimated slightly differently ($\beta_{\text{unter}} = -0.0574$, 95% CI $[-0.1162, -0.0002]$). If not noted otherwise, I collapsed data from *über* and *unter* trials in all following analyses.

I could also replicate a second finding from the literature: LOs that are located on the side of the grazing line that corresponds to the to-be-rated preposition are rated higher than LOs that are located on the non-corresponding side of the grazing line (Regier & Carlson, 2001, exps. 5 & 6). The grazing line is the imaginary horizontal line that touches the top points of the RO (for superior prepositions; for inferior prepositions, the grazing line touches the bottom of the RO). I contrasted two subsets of the rating data to test for the effect of the grazing line (see Figure 4.8): Ratings for LOs on the corresponding side of the grazing line (i.e., above

the grazing line for *über, above*, rows R1–R3, and below the grazing line for *unter, below*, rows R8–R10) against ratings for LOs on the grazing line or on the non-corresponding side of the grazing line (rows R4–R7). I specified an ordinal regression model that predicts rating from this two-level predictor (corresponding vs. non-corresponding side of grazing line).

To integrate the quantitative findings from Regier and Carlson (2001, exps. 5 & 6), I set the mean of the Gaussian prior distribution of the slope parameter to $\mu = 3.7$ (see also Table 3.5 on page 64). I chose a relatively large standard deviation $\sigma = 3.0$ to account for methodological differences (Regier & Carlson, 2001, treated their ordinal data as metric and conducted their experiments in English). The posterior distribution for this regression model confirms the grazing line effect for the data presented here: Participants rated LOs on the corresponding side of the grazing line (rows R1–R3 & rows R8–R10) credibly higher than LOs on the non-corresponding side of the grazing line (rows R4–R7; $\beta_{\text{corresponding}} = 3.49$, 95% CI [3.34, 3.65]). Figure 4.10b depicts this effect: For roughly 80% of LOs on the non-corresponding side, participants picked ratings 1 or 2.

The grazing line effect is also clearly visible in the data visualization that I created separately for each RO (Figures 4.11, 4.12, 4.14, and 4.15). In these visualizations, all individual ratings are plotted as color-coded rhombi on top of each other close to the location of the rated LO. The brighter the rhombus, the smaller the rating. Ratings for LOs in rows R4–R7 stand out as being bright (i.e., low) in comparison to all other LOs.

Acceptability Ratings: Asymmetrical ROs

The analysis of the ratings for LOs around the asymmetrical ROs (see Figures 4.11 and 4.12 for a visualization) could also be interpreted as an attempt to replicate the effect of the center-of-mass orientation. Regier and Carlson (2001, exps. 1–3) provide evidence that the center-of-mass orientation affects acceptability ratings. However, apart from their fourth experiment, they only used symmetrical, rectangular ROs. While their fourth experiment was explicitly designed to contrast the center-of-mass with the midpoint of an asymmetrical, triangular RO, it tested only 4 LOs around 2 ROs (8 LOs in total, see Section 3.2.4 for stimuli and data). My stimuli extend the number of LOs around asymmetrical ROs to 224 LOs (28 LOs \times 4 ROs \times 2 prepositions). However, different than the upright triangle in Regier and Carlson's experiment 4, my ROs faced the LOs only with a flat surface (see flipped versions of the L and mL ROs in Figure 4.12).

To test for the effect of the center-of-mass orientation, I contrasted ratings for LOs with equal average center-of-mass orientation. To do so, I created two data subsets (corresponding to the contrast sets in the PSP analysis): A “mass” subset with ratings for LOs that are located directly

Higher ratings for LOs on the side of the grazing line that corresponds to the used preposition (vs. LOs on the non-corresponding side).

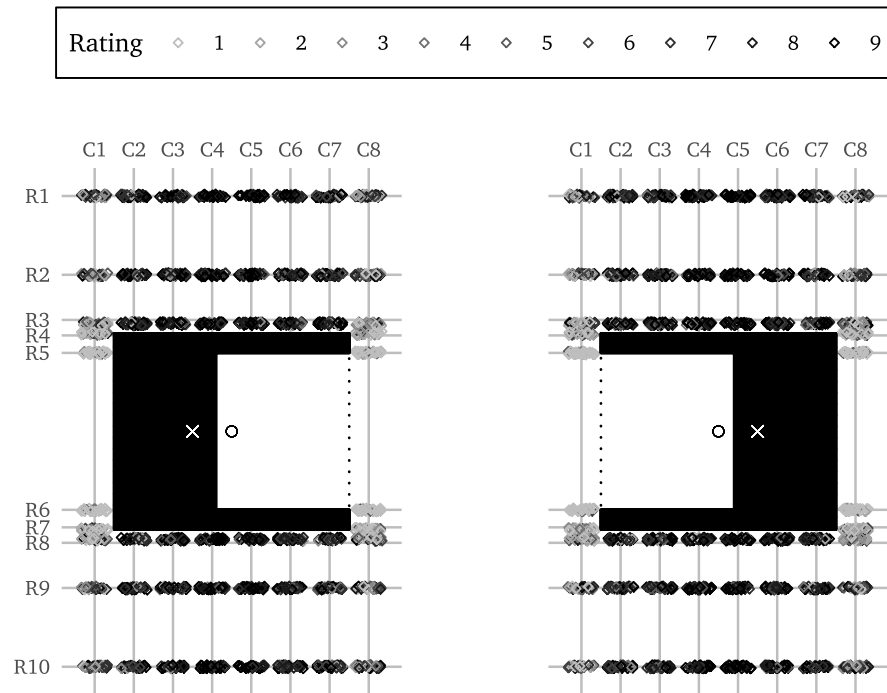


Figure 4.11: Individual *über* (*above*) and *unter* (*below*) acceptability ratings for LOs (not depicted) around the asymmetrical C and mC ROs. Individual acceptability ratings are color-coded (the darker the color, the higher the rating) and plotted near the location of the corresponding LO (to avoid overplotting). LOs (not shown in the visualization) were black circles with a 10-pixel diameter and placed at the intersection of the grid lines. LOs in rows R1–R5 were presented with *über* (*above*), LOs in rows R6–R10 were presented with *unter* (*below*). Only one RO and one LO was visible at a time. For each RO: Dashed line is the bounding box, \times is the center-of-mass, \circ is the center-of-object. Neither of the centers nor the bounding box were visible to the participants. Image copyright: See Appendix E.

above the mass side of the asymmetrical ROs (columns C2 & C3 for ROs C and L; columns C6 & C7 for ROs mC and mL, see Figures 4.11 and 4.12) and a “cavity” subset with ratings for LOs above the cavity of the asymmetrical ROs (columns C4 & C5 for all asymmetrical ROs). Based on center-of-mass orientation only, I would expect no difference in ratings for these two subsets. This is because, on average, the center-of-mass orientation is equal for both subsets. In addition, the proximal orientation is constant for these LOs.

I specified a Bayesian regression model predicting rating from membership in either subset. I explicated the prior expectation of finding no effect as the following prior distribution for the slope parameter: a Gaussian distribution centered at $\mu = 0.0$ with a narrow spread of $\sigma = 0.1$. Despite this prior, the regression model reveals credibly lower ratings for LOs in the “mass” subset compared to LOs in the “cavity” subset

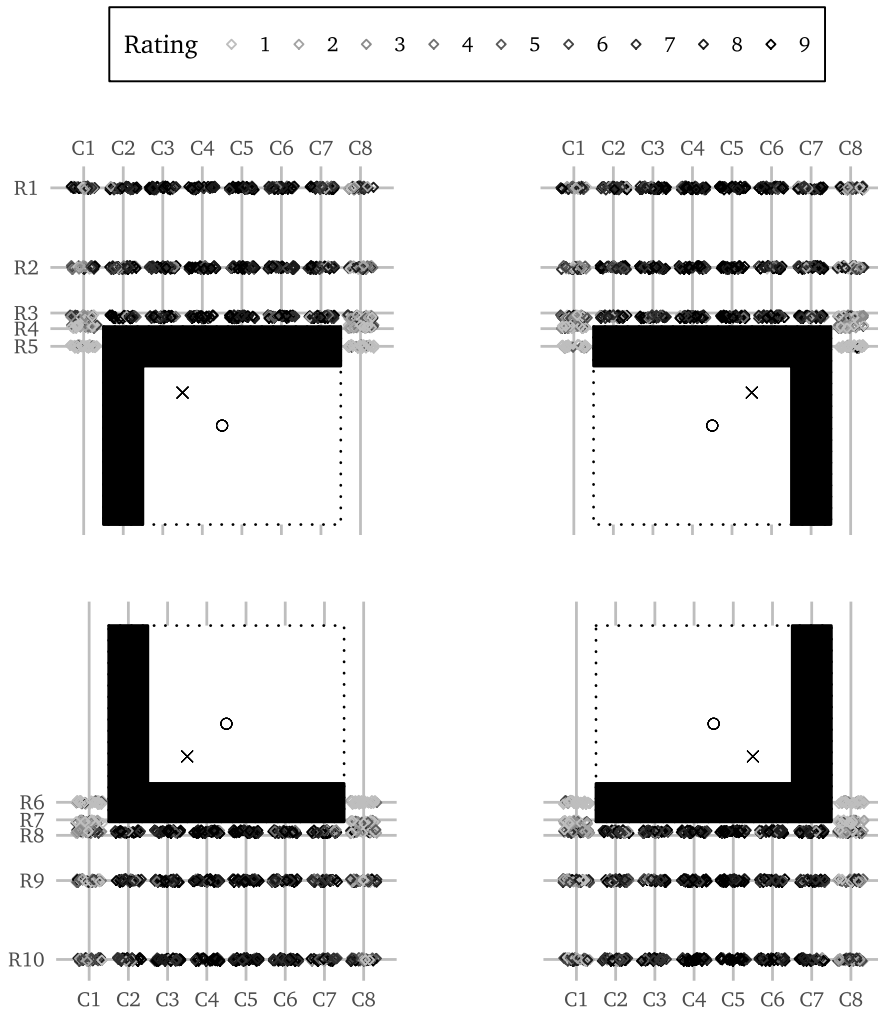


Figure 4.12: Individual *über* (*above*) and *unter* (*below*) acceptability ratings for LOs (not depicted) around the asymmetrical L and mL ROs. Individual acceptability ratings are color-coded (the darker the color, the higher the rating) and plotted near the location of the corresponding LO (to avoid overplotting). LOs (not shown in the visualization) were black circles with a 10-pixel diameter and placed at the intersection of the grid lines. LOs in rows R1–R5 were presented with *über* (*above*), LOs in rows R6–R10 were presented with *unter* (*below*). Only one RO and one LO was visible at a time. For each RO: Dashed line is the bounding box, \times is the center-of-mass, \circ is the center-of-object. Neither of the centers nor the bounding box were visible to the participants. Image copyright: See Appendix E.

($\beta_{\text{mass}} = -0.84$, 95% CI $[-0.97, -0.71]$). I recomputed the same model with the uninformative default prior provided by the brms package. This model shows the same qualitative effect. Quantitatively, it even reveals a greater effect size ($\beta_{\text{mass}} = -1.46$, 95% CI $[-1.63, -1.29]$).

Higher ratings for LOs in the “cavity” subset (compared to LOs in the “mass” subset).

To assess which of the two models better fits the data, I computed the LOO criterion (Vehtari et al., 2017). The LOO measure favored the model with the uninformative prior over the model with the informative prior (lower LOO value for the default-prior-model, 5 631.44, compared to the null-effect-prior model, 5 680.88). Figure 4.13 plots the prediction of the model with the default prior alongside the empirical rating distribution. The plots show that the regression model accounts well for the data. Moreover, it is visible that participants chose rating 9 more often for LOs in the “cavity” subset than for LOs in the “mass” subset. In contrast, for LOs in the “mass” subset they picked rating 7 & 8 more often than for LOs in the “cavity” subset.

One possible explanation for this rating pattern is that people do not base their acceptability judgment on the center-of-mass of the RO (marked with \times in Figures 4.11 and 4.12) but rather use the center-of-object (depicted as \circ in Figures 4.11 and 4.12). Here, the center-of-object is the center of the ‘bounding box’ (BB) of the RO. The BB of an RO is the smallest rectangle containing all points of the RO. For the rectangular ROs, the BB coincides with the RO. For the asymmetrical ROs, the BB also includes the cavities of the objects (see dashed rectangles in Figures 4.11 and 4.12). More precisely, the center-of-object is defined as

The center-of-object is the center of the bounding box of the RO.

$$\text{CoO}(x, y) = \left(\text{RO}_{x_0} + \frac{\text{RO}_{\text{width}}}{2}, \text{RO}_{y_0} + \frac{\text{RO}_{\text{height}}}{2} \right) \quad (4.1)$$

Here, RO_{x_0} is the leftmost point of RO’s BB and RO_{y_0} is the point with the lowest y-coordinate (y-axis increasing from bottom to top). For rectangular ROs, the center-of-mass and the center-of-object coincide. Given that LOs in the “cavity” subset are more central with respect to the center-of-object than LOs in the “mass” subset, the lower ratings for LOs in the “mass” subset could be explained in terms of higher center-of-object orientations.

I computed four further Bayesian regression models to test this hypothesis. The first two models predict rating as a function of the RO-side at which the LO was located. First, I split the LOs in two subsets that were either left (columns C1–C4) or right (C5–C8) from the center-of-object. On average, the LOs in these subsets have the same center-of-object orientation. Thus, if participants consider the center-of-object orientation, then I would expect no credible rating differences in these subsets. This prediction is confirmed by the statistical model ($\beta_{\text{right}} = 0.05$, 95% CI $[-0.03, 0.13]$).

The second regression model applied another subsetting of the data to investigate the influence of the asymmetrical mass distribution. To this end, each LO was classified as either being on the side where the center-of-mass of the RO was located or on the other side. For example, the center-of-mass-side-subset for the L RO consists of LOs in columns C1–C4 whereas the center-of-mass-side-subset for the mL RO consists of LOs in columns C5–C8 (see Figure 4.12). This model

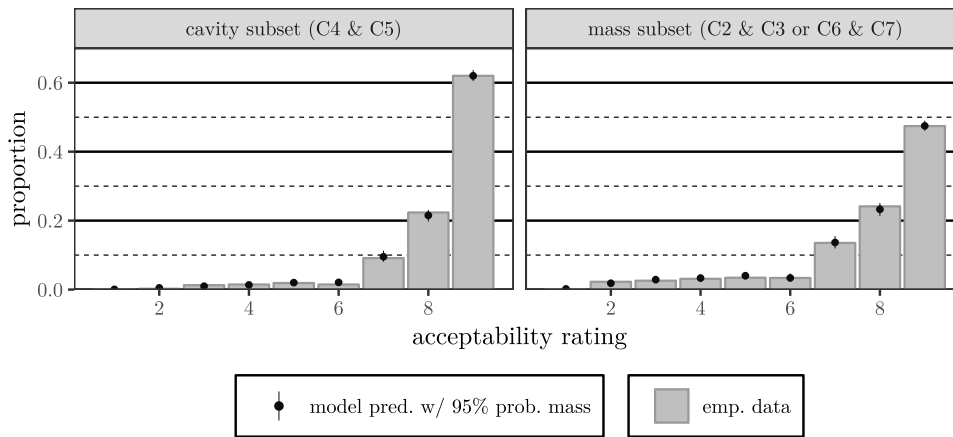


Figure 4.13: Empirical rating distributions and fit of Bayesian ordinal regression model (computed with 100 samples from the posterior distribution) contrasting ratings for the two subsets “cavity” (columns C4 & C5, all asymmetrical ROs) and “mass” (columns C2 & C3 for ROs C and L; columns C6 & C7 for ROs mC and mL). Bayesian regression model was computed with brms’s default prior. Image copyright: See Appendix E.

reveals a small but credible effect of the asymmetrical mass distribution. LOs that were located on the same side as the center-of-mass of the RO received a higher rating compared to LOs on the cavity side of the RO ($\beta_{\text{CoMside}} = 0.10$, 95% CI [0.02, 0.18]). In terms of ratings, this model estimates a 2% higher probability for ratings 8 and 9 for LOs on the center-of-mass side compared to LOs on the cavity side. Thus, the location of the center-of-mass seems to affect acceptability ratings. However, the effect is smaller than expected if one assumes that people only consider the center-of-mass orientation.

To directly contrast the center-of-mass and the center-of-object orientation, I specified two Bayesian regression models that predicted rating as a function of either the center-of-mass orientation or the center-of-object orientation (in radian notation and centered). Both models reveal a credible influence of each orientation ($\beta_{\text{CoM}} = -4.58$, 95% CI [-4.73, -4.42]; $\beta_{\text{CoO}} = -7.24$, 95% CI [-7.46, -7.02]). Crucially, however, the center-of-mass effect is smaller than the center-of-object effect. Moreover, the center-of-object model fits the data better than the center-of-mass model (as measured via the LOO criterion: 23 235.51 for the center-of-mass model, 21 175.10 for the center-of-object model; lower LOO is better). In addition, a pairwise model comparison using the Bayes factor (see e.g., Mulder & Wagenmakers, 2016) favored the center-of-object model over the center-of-mass model.

In summary, my analyses suggest that the center-of-object orientation has a greater influence on acceptability ratings than the center-of-mass orientation. This conflicts with the assumption of the importance of the center-of-mass orientation in the AVS and the rAVS_{w-comb} model. There-

People seem to base their ratings on the center-of-object instead of on the center-of-mass.

fore, I present two model modifications in Section 5.1 that integrate the center-of-object orientation instead of the center-of-mass orientation in their computations.

Acceptability Ratings: Relative Distance

So far, I only analyzed one half of the collected rating data. The second half of the rating data comes from LOs around four rectangular ROs. I designed these stimuli to test for a potential effect of relative distance on acceptability judgments. Roughly, relative distance is defined as absolute distance divided by the dimensions of the RO (see Equation 3.5 on page 37). Thus, keeping the LO position constant but increasing the height of the RO reduces the relative distance of the LO to the RO. This is why I used four rectangular ROs with different heights. Figures 4.14 and 4.15 visualize the individual ratings for these stimuli.

To analyze these data, I started with asking whether the acceptability ratings differ for the four ROs. To this end, I specified a Bayesian regression model that predicts rating from RO (thin, thick, square, or tall rectangle). Since both the AVS and the $rAVS_{w-comb}$ model partly predicted higher ratings for LOs above taller rectangles compared to LOs above thinner rectangles (see PSP results, Section 4.1.3), I specified Gaussian prior distributions with $\mu = 0.5$ for each RO's regression parameter. These prior distributions tendentially support the existence of the predicted effect. However, given that this is the first study testing for the effect of relative distance, I chose relatively broad prior distributions with $\sigma = 1.5$. These distributions also allow for a null or a reversed effect. Indeed, despite the supporting prior, the regression model reveals no credible difference in rating distributions between the different ROs (see Figure 4.16 for empirical rating distributions and model fits; $\beta_{thick} = 0.01$, 95% CI $[-0.11, 0.12]$; $\beta_{square} = 0.02$, 95% CI $[-0.09, 0.14]$; $\beta_{tall} = 0.04$, 95% CI $[-0.08, 0.15]$; thin rectangle was the intercept of the regression model). The same regression model with brms's uninformative default prior results in almost the same estimates ($\beta_{thick} = 0.00$, 95% CI $[-0.11, 0.12]$; $\beta_{square} = 0.02$, 95% CI $[-0.10, 0.14]$; $\beta_{tall} = 0.04$, 95% CI $[-0.08, 0.15]$).

*Rating patterns are
not affected by
rectangle height.*

This finding goes against the qualitative model predictions from both the AVS and the $rAVS_{w-comb}$ model that LOs above taller rectangles should receive higher ratings compared to LOs above thinner rectangles. However, both cognitive models also allow the no-difference case that exists in the empirical data (see PSP results in Figure 4.6 on page 84). Considering the proposed relative distance mechanism of the $rAVS_{w-comb}$ model, the empirical data are even more interesting. The $rAVS_{w-comb}$ model proposes that relative distance affects the way people weight the influences of the proximal orientation and the center-of-mass orientation on their acceptability judgment: With low relative distance, the $rAVS_{w-comb}$ model considers the proximal orientation as more important than the center-of-mass orientation whereas with high relative

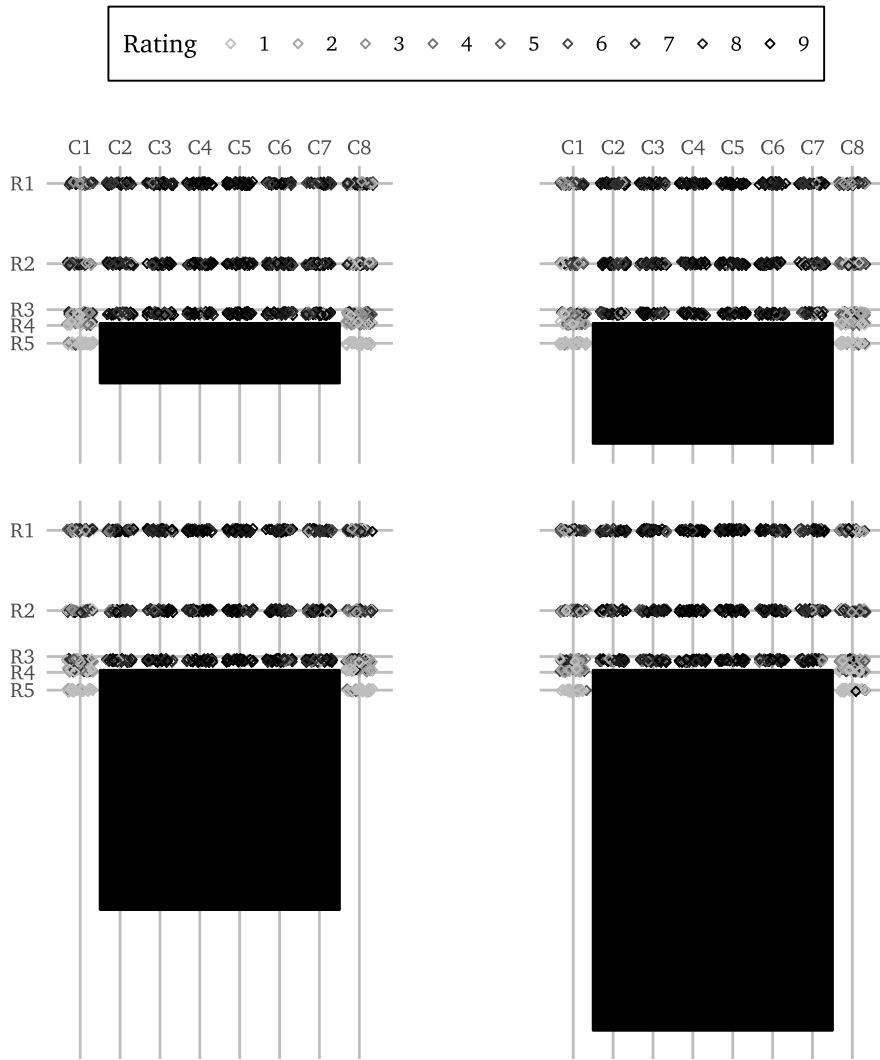


Figure 4.14: Individual *über* (*above*) acceptability ratings for LOs (not depicted) above the thin, the thick, the square, and the tall rectangle. Individual acceptability ratings are color-coded (the darker the color, the higher the rating) and plotted near the location of the corresponding LO (to avoid overplotting). LOs (not shown in the visualization) were black circles with a 10-pixel diameter and placed at the intersection of the grid lines. Only one RO and one LO was visible at a time. Image copyright: See Appendix E.

distance the center-of-mass orientation becomes more important than the proximal orientation. For the tested stimuli, the proximal orientation does not change across the different rectangular ROs. However, the center-of-mass orientation decreases with increasing RO height. This is because the center-of-mass of a taller rectangle is located lower than the center-of-mass of a thinner rectangle (if both rectangles are aligned at their tops). Thus, according to the known influence of center-of-mass orientation, one would expect higher ratings for taller rectangles – without considering the factor relative distance at all. Apparently, relative

The height of an RO seems to interact with the effect of the center-of-mass orientation.

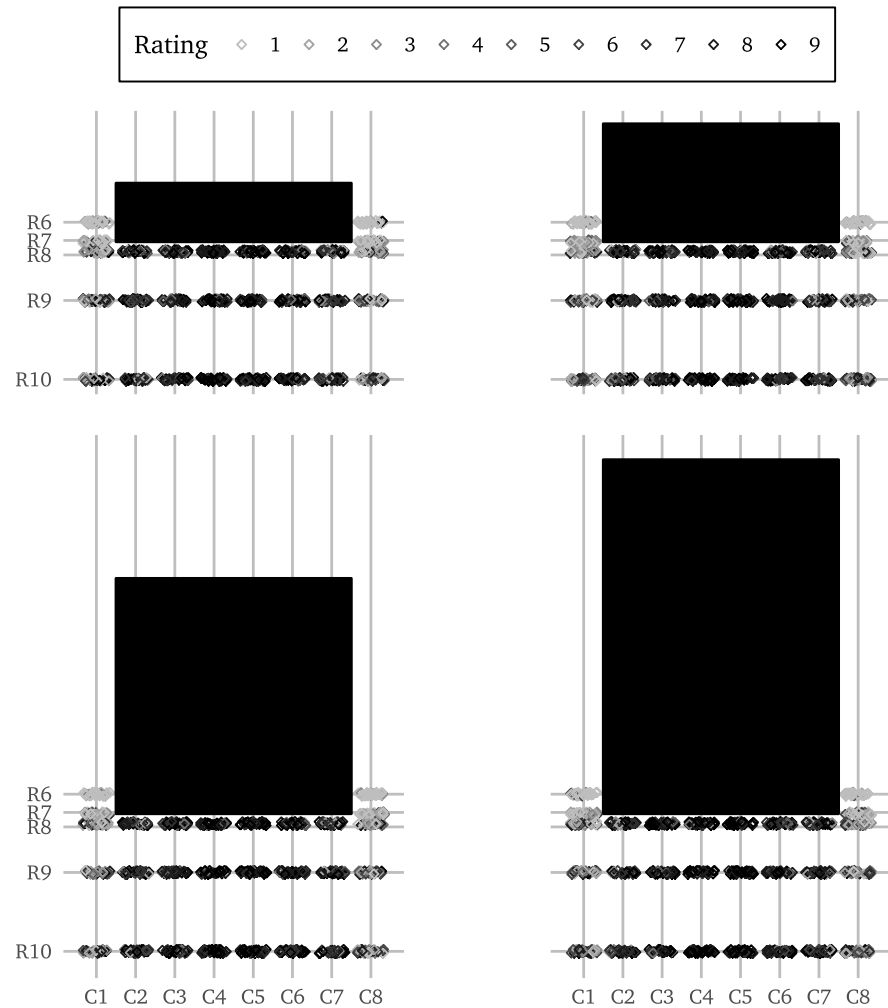


Figure 4.15: Individual *unter* (*below*) acceptability ratings for LOs (not depicted) below the thin, the thick, the square, and the tall rectangle. Individual acceptability ratings are color-coded (the darker the color, the higher the rating) and plotted near the location of the corresponding LO (to avoid overplotting). LOs (not shown in the visualization) were black circles with a 10-pixel diameter and placed at the intersection of the grid lines. Only one RO and one LO was visible at a time. Image copyright: See Appendix E.

distance somehow affects the processing of center-of-mass orientation as part of generating an acceptability judgment.

Generally speaking, this hypothesis (relative distance affects center-of-mass and proximal orientation) is in line with the $rAVS_{w-comb}$ model. With the following analysis, I investigated whether the empirical data speak to $rAVS_{w-comb}$'s particular mechanism. To this end, I specified a Bayesian regression model that uses the predictors relative distance,

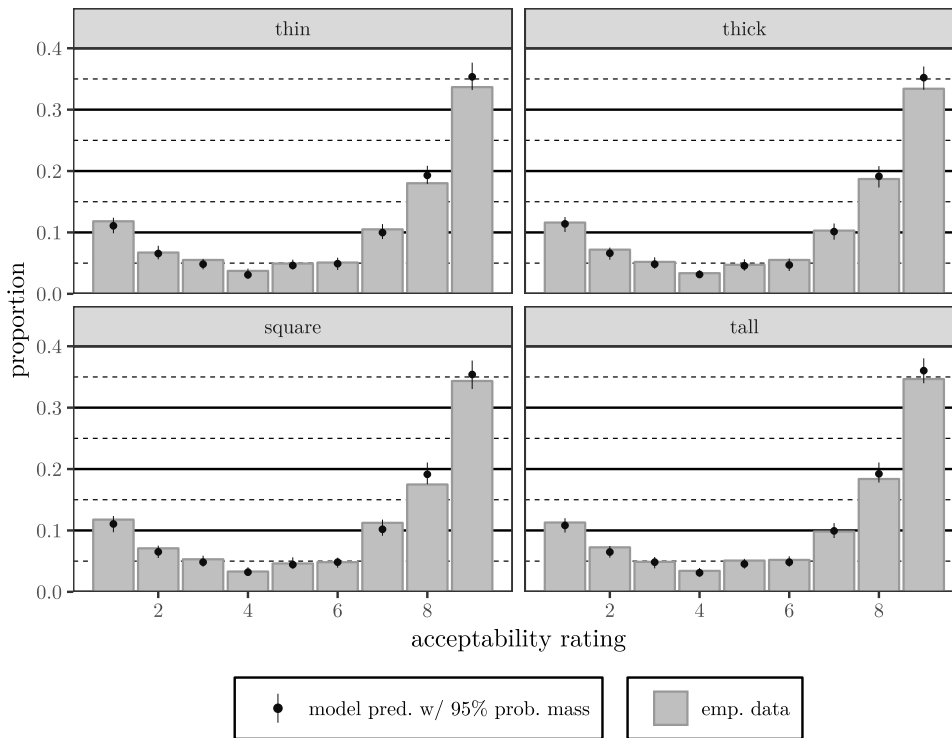


Figure 4.16: Empirical rating distributions and fit of Bayesian ordinal regression model (computed with 100 samples from the posterior distribution) contrasting ratings for the four rectangular ROs. Bayesian regression model was computed with prior distributions supporting higher ratings for taller rectangles. Image copyright: See Appendix E.

center-of-mass orientation⁶, and proximal orientation to predict the outcome rating. I centered all predictors and used radian notation for the two orientational predictors. For computing the predictor relative distance for each LO, I applied $rAVS_{w-comb}$'s definition (see Equation 3.5 on page 37). Furthermore, I allowed full interactions between all predictors. For comparison, I computed all simpler competitor models by removing interactions or predictors from the model. As revealed by the LOO method, the most complex model (presented here) fits the data best (lowest LOO). In addition, pairwise model comparisons using Bayes factors (see e.g., Mulder & Wagenmakers, 2016) favored the most complex model over all simpler models.

Figure 4.17 shows two perspectives on this complex model. Figure 4.17a plots the estimated effect of proximal orientation on acceptability rating and Figure 4.17b plots the effect of center-of-mass orientation on acceptability rating. Note that for ease of visualization the plots treat the outcome variable as metric which is an incorrect assump-

⁶ Note that for rectangular ROs the center-of-mass orientation coincides with the center-of-object orientation.

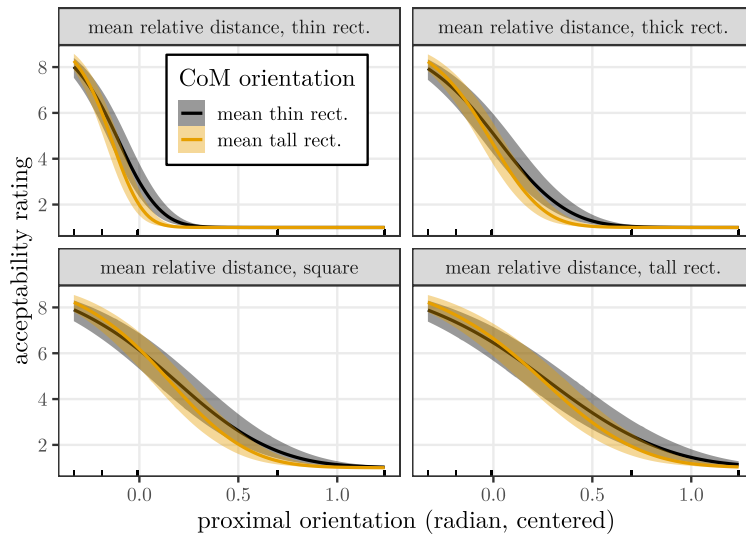
tion. This is not how the outcome variable is actually handled by the ordinal regression model. Also for ease of visualization, I had to keep two of the three predictors at constant values for the plots. I kept the predictor relative distance constant at its mean values for LOs around each rectangle (subplots in Figure 4.17a, colored lines in Figure 4.17b). In Figure 4.17a, the predictor center-of-mass orientation is constant on its mean value for LOs around the thin or the tall rectangle (colored lines). In Figure 4.17b, the predictor proximal orientation is constant on non-deviating orientation (LOs directly above the RO, left subplot) and the mean of all deviating proximal orientations (right subplot).

What do these plots tell us about how relative distance influences the effect of either center-of-mass orientation or proximal orientation on acceptability rating? Considering the proximal orientation effect first (Figure 4.17a), the model shows that higher proximal orientation correlates with lower acceptability ratings (negative slopes in all subplots). This is in line with the known effect of proximal orientation. Interestingly, however, relative distance modulates this effect: With smaller relative distance (i.e., for larger rectangles) the strength of the proximal orientation effect shrinks. This is evident from comparing the steepnesses of the slopes in the four subplots of Figure 4.17a: The smaller the relative distance, the less steep is the slope. Different values of center-of-mass orientation also affect the steepness of the slope, although to a different degree. For high values of center-of-mass orientation (i.e., for thinner rectangles; black lines in Figure 4.17a), the effect of proximal orientation is slightly less pronounced (i.e., the slope is less steep) than for low values of center-of-mass orientation (i.e., for taller rectangles; yellow lines in Figure 4.17a). However, this modulation should be treated as a small trend, given that the 95% CIs of the black and yellow lines overlap considerably almost everywhere.

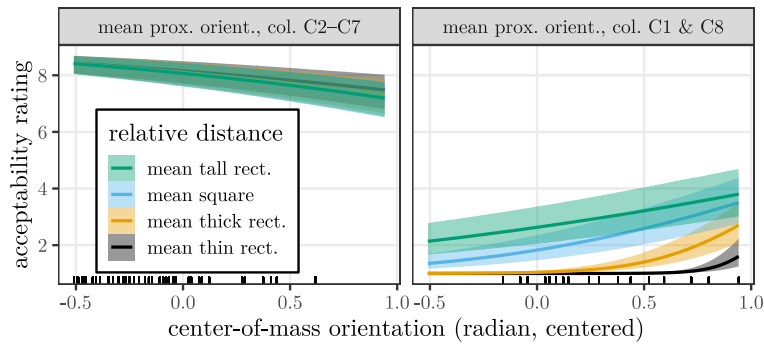
How is the effect of the center-of-mass orientation modulated by different values of relative distance? Figure 4.17b plots the center-of-mass orientation effect on acceptability rating. For non-deviating proximal orientation (i.e., for LOs directly above the RO, columns C2–C7, left subplot in Figure 4.17b), center-of-mass orientation affects acceptability ratings as expected: The higher the center-of-mass orientation, the lower the rating (negative slope). The four different values of relative distance do not change this observation: the 95% CIs of all colored lines overlap almost entirely. The right subplot of Figure 4.17b (depicting the model estimates for higher values of proximal orientation) surprisingly shows a reversed effect: Here, higher center-of-mass orientation correlates with higher ratings (positive slopes). In addition, relative distance modulates this reversed effect: For high relative distance (for thinner rectangles, black and yellow lines in Figure 4.17b), the reversed effect is more pronounced (the slopes are steeper) than for low relative distance (for taller rectangles, blue and green lines).

The lower the relative distance, the less pronounced is the effect of proximal orientation.

Relative distance modulates a reversed effect of center-of-mass orientation.



(a) Effect of proximal orientation on acceptability rating with constant values for predictors relative distance (subplots) and center-of-mass orientation (colored lines).



(b) Effect of center-of-mass orientation on acceptability rating with constant values for predictors proximal orientation (subplots) and relative distance (colored lines).

Figure 4.17: Visualization of effects of (a) proximal orientation and (b) center-of-mass orientation on acceptability rating as estimated by a Bayesian regression model with these two predictors plus relative distance. Plots treat outcome variable as metric (for visualization purposes) which is not how the ordinal regression model deals with the data. Predictors not on the x-axis were kept constant on meaningful values: Relative distance is constant on its mean values for LOs around each of the four ROs, center-of-mass orientation is constant on mean values for LOs around the thin or tall rectangle, proximal orientation is constant on non-deviating orientation (columns C2–C7) and the mean value of deviating proximal orientation (columns C1 & C8). Little black bars on the x-axis denote actually tested data points. Shaded areas denote 95% CIs. Image copyright: See Appendix E.

Taken together, my analysis confirms the general proposal from the $rAVS_{w-comb}$ model that relative distance affects the influence of center-of-mass orientation and proximal orientation on acceptability ratings. However, the specific $rAVS_{w-comb}$ mechanism is not confirmed. The $rAVS_{w-comb}$ model assumes that (i) higher relative distance leads to a higher influence of the center-of-mass orientation compared to the proximal orientation and correspondingly that (ii) lower relative distance leads to a higher influence of proximal orientation compared to center-of-mass orientation. In contrast, the empirical data suggest that (i) higher relative distance strengthens a *reversed* effect of center-of-mass orientation and that (ii) lower relative distance weakens the effect of proximal orientation.

4.2.2 Results: Eye Movements

Due to my study design that presented written sentences before the spatial configurations (in order to stay close to earlier studies, e.g., Hayward & Tarr, 1995; Logan & Sadler, 1996; Regier & Carlson, 2001), I cannot interpret the recorded eye movements in terms of time-locked linguistic processing of the unfolding spatial utterance. However, the gaze patterns still reflect deployment of overt visual attention during spatial relation processing. This is why they are still interesting for my research question. More specifically, I analyzed the fixation data in order to answer two main questions: First, do people preferably look at the attentional focus point as defined by the AVS model (also an important point in the $rAVS_{w-comb}$ model)? Second, do the gaze patterns reflect the mass distribution of the asymmetrical ROs – although participants rated LOs around these ROs as if they had no cavities?

Before I answer these questions, I introduce the data set more closely. The center-of-mass of every RO was located at the center of the screen. The spatial preposition was approximately in the middle of the sentence, which was also close to the center of the screen (see Figure 4.9 on page 88). Some participants reported that after a few trials they did not attend to the whole sentence anymore, because they figured out that the experiment consisted of only two sentences and only the spatial preposition was relevant for the task. Since I did not use a fixation cross elsewhere on the screen, it is very likely that many participants fixated near the center of the screen after they confirmed the sentence, i.e., when the RO appeared. Indeed, more than 46% of the fixations that started in the first 150 ms after the RO was shown were inside a 100 pixel (ca. 2.02 degrees of visual angle) wide square around the center of the screen (i.e., these fixations had at most 50 pixel, ca. 1.01 degree of visual angle, distance in either direction to the center of the screen).

Given that the center-of-mass of the RO is a point I am especially interested in and the planning of a saccade takes approximately 200 ms (Matin, Shao, & Boff, 1993, cited in Tanenhaus, Spivey Knowlton, Eber-

hard, & Sedivy, 1995), I only analyzed fixations that started more than 150 ms after the onset of the spatial configuration. Of these 53 718 fixations (mean number of fixations per subject: 1 579.94, standard deviation: 688.76; mean number of fixations per subject and trial: 3.53, standard deviation 2.82) roughly half (ca. 46%) landed close to the LO (no more than 45 pixels in x or y direction) and ca. 21% landed inside the bounding box of the RO. The bounding box (BB) of an RO is the smallest rectangle containing all points of the RO. For the rectangular ROs, the BB coincides with the RO. For the asymmetrical ROs, the BB also includes the cavities of the objects (see dashed rectangles in Figures 4.11 and 4.12). Given that the LOs in rows R3–R8 were close to the RO (15 pixels) and the accuracy of the eye tracker is of the same magnitude, some fixations in these trials were counted as both: close to the LO and inside the bounding box of the RO.

Eye Movements: Fixations to Hypothetical Focus Point

Using the fixations inside the BB, I answer the first question whether people preferably fixate AVS's attentional focus F. The attentional focus F in the AVS model is defined to be the point on top of the RO that is vertically aligned with the LO (for superior prepositions; for inferior prepositions it is the corresponding point on the bottom of the RO). If the LO is not in the region directly above the RO, AVS's focal point is the closest point on top of the RO (i.e., either the top-right or the top-left corner of the RO). This point F plays also an important role in the rAVS_{w-comb} model. However, apart from Carlson et al. (2006), I am not aware of any eye tracking study that explicitly assessed whether this point F is actually a point that people preferably fixate.

To investigate this issue, I first plotted all fixations in the BBs of the ROs as heatmaps in Figures 4.18 and 4.19. Figure 4.18 plots the number of fixations in absolute coordinates (pixels), Figure 4.19 plots the same fixations in relative coordinates, i.e., normalized with respect to the dimensions of the BB of the RO. These heatmaps show that participants primarily fixated the top of the RO for *über* (*above*) and the bottom of the RO for *unter* (*below*). This partly confirms the vertical location of the assumed point F. However, since I did not include trials that tested *über* (*above*) with LOs that were located below the RO (nor *unter*, *below*, trials with LOs above the RO), it remains unclear whether people fixated the top/bottom because of the location of the LO or because of the preposition they had to rate. Future studies should untangle the visual influence (LO placement) from the linguistic influence (preposition) by testing clear mismatches of LO placement and preposition.

While the vertical component of point F remains to be tested, there is evidence that fixations correlate with the horizontal component of point F. In Figure 4.20, I plotted a heatmap of relative fixation locations in the BB by the column of the used LO (C1–C8). These heatmaps show that the horizontal location of the LO affects the horizontal fixation location

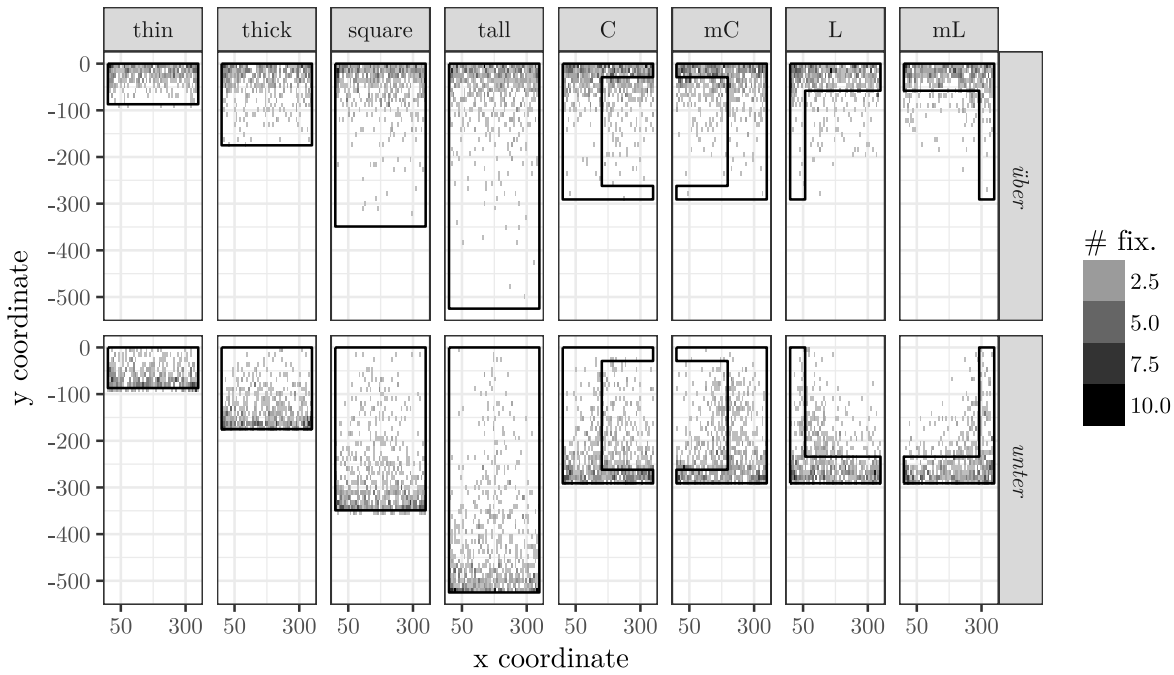


Figure 4.18: Heatmap visualizations depicting the number of fixations inside BBs of ROs, separated by RO and preposition. Coordinates are in pixels, starting to count from top left of each RO. Computed with 50×50 bins.

The horizontal location of the LO predicts the horizontal location of fixations in the BB of the RO.

on the RO: LOs placed above-left of the RO (C1–C3) correlate with more fixations on the left part of the RO while LOs placed above-right of the RO (C6–C8) correlate with more fixations on the right part of the RO. Finally, centrally placed LOs (C4 & C5) result in more fixations on the central part of the RO. This is true for both prepositions.

As a further test of this visual evidence, I specified a Bayesian regression model that predicted horizontal relative fixation inside the BB (i.e., relative x coordinate as plotted in Figure 4.20) from the x-coordinate of the LO (in pixels). This model shows a credible influence of horizontal LO location on horizontal fixation location: The more right the LO was placed, the more right landed the fixation inside the BB ($\beta_{LO_x} = 0.241$, 95% CI[0.237, 0.245]). To make sure that fixations aimed for the RO and not the LO, I excluded data from trials with LOs very close to the RO (rows R3–R8). This regression model with a smaller data subset provides the same qualitative results with a slightly different regression coefficient ($\beta_{LO_x} = 0.203$, 95% CI [0.196, 0.210]). Taken together, the data support the importance of the point F (as assumed in the AVS and the rAVS_{w-comb} models) by showing that it is indeed a point that attracts fixations.

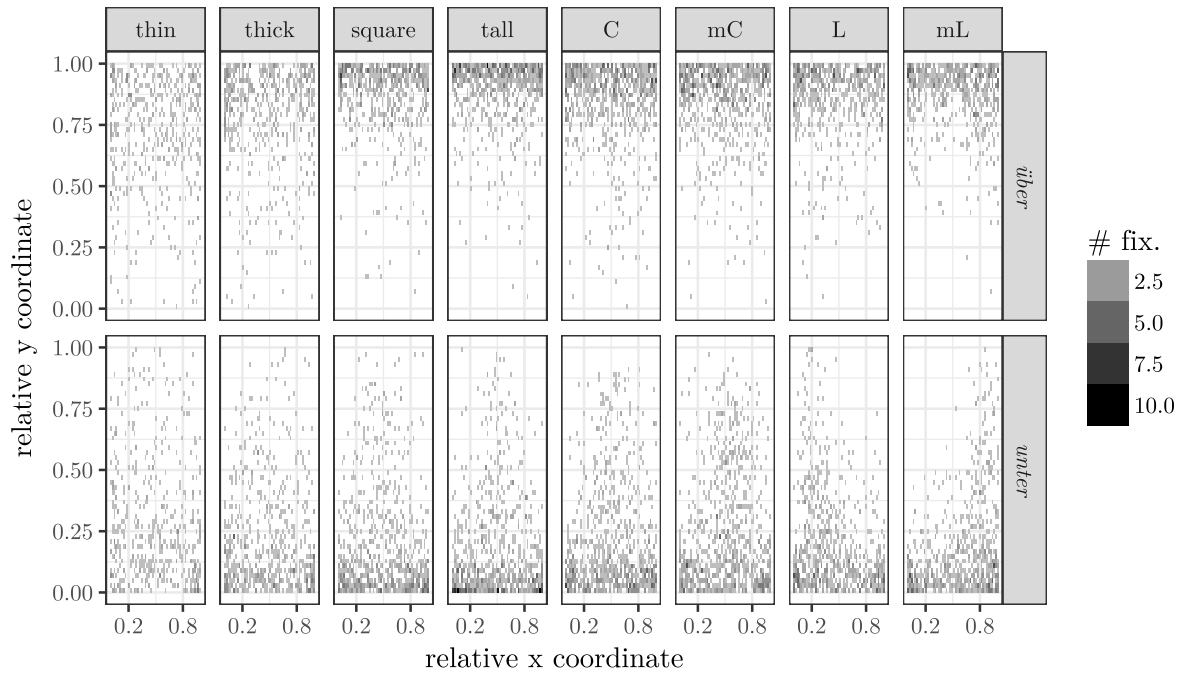


Figure 4.19: Heatmap visualizations depicting the number of fixations inside BBs of ROs, separated by RO and preposition. Coordinates are normalized by the dimensions of each BB such that they are relative to each BB. Computed with 50×50 bins.

Eye Movements: Asymmetrical ROs

The second question I wanted to answer with the eye movement data concerns the processing of the asymmetrical ROs. The rating pattern for the asymmetrical ROs (center-of-object appears to be more important than center-of-mass) suggests that participants processed the asymmetrical ROs as if they were rectangular ROs. I was interested whether the gaze patterns mirror this finding or whether they potentially reflect the asymmetrical mass distributions of the ROs.

A first answer to this question is the number of looks to either the center-of-mass or the center-of-object. The right part of Table 4.1 provides the number of fixations close to either center (i.e., no more than 25 pixel in x or y direction from a center). Note that for the rectangular ROs, the center-of-mass and center-of-object coincide. Here, Table 4.1 provides the overall number of fixations to the single center of the RO in both columns. Considering the looks to the centers, it is evident that no center served as an attractor for people's fixations: Out of all 11 335 fixations that landed inside the BBs (see first column of Table 4.1 for RO-wise counts), only 315 fixations (ca. 2.8%) landed close to a center (rectangular ROs: 115 fixations, ca. 2.2%; asymmetrical ROs: 200 fixations, ca. 3.2%). Interestingly, though, is the fact that the centers of the asymmetrical ROs attracted more fixations than most of the centers of

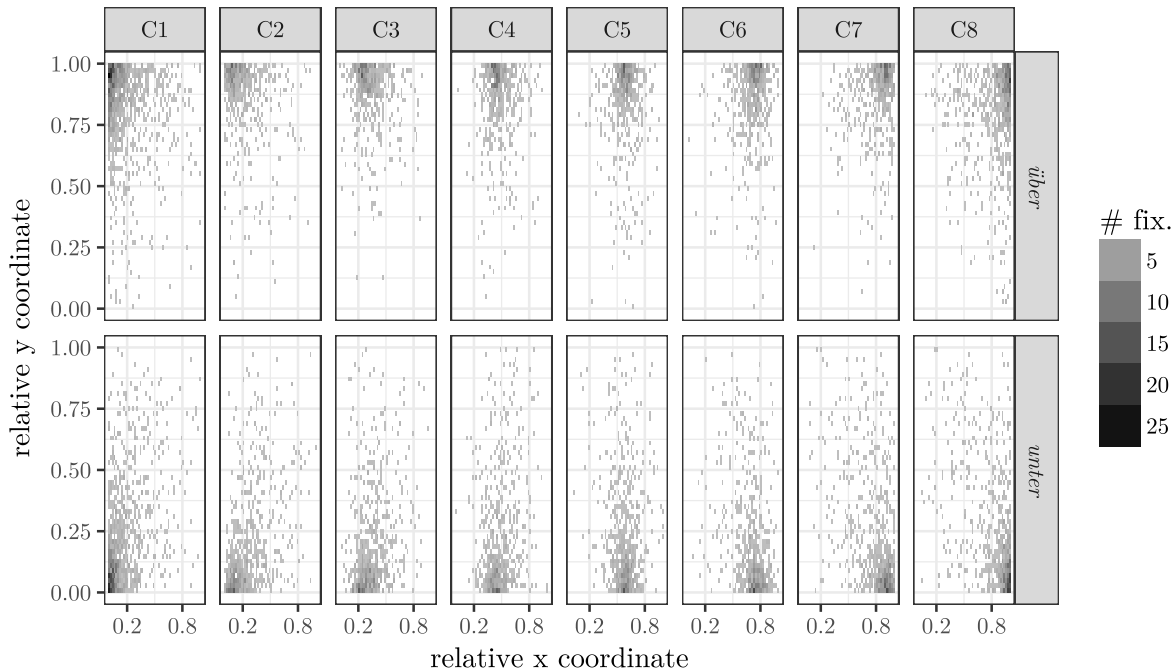


Figure 4.20: Heatmap visualizations depicting the number of fixations inside BBs of ROs, separated by the column of the LO and preposition. Coordinates are normalized by the dimensions of each BB such that they are relative to each BB. Computed with 50×50 bins. Image copyright: See Appendix E.

the rectangular ROs. Another interesting trend is that taller rectangles have more total fixations than thinner rectangles.⁷

In terms of asymmetrical gaze patterns, the counts in Table 4.1 suggest that participants fixated the center-of-mass more than the center-of-object for the L and mL RO, while they slightly preferred the center-of-object for the C and mC ROs. However, at least for the L and mL ROs, this conclusion is flawed because their centers are not on the same vertical level (see Figure 4.12): The center-of-mass is shifted in the direction of the top or bottom of the RO. Given that earlier analyses suggested that participants preferred to fixate the top/bottom of the RO, a higher number of fixations on the closer center-of-mass compared

⁷ To test a potential linking hypothesis that maps number of fixations to processing difficulty (as measured via reaction time), I specified a Bayesian regression model that predicts reaction time from RO (for more RT analyses, see page 110). Partly supporting my hypothesis, this model reveals credibly longer reaction times for the tall rectangle and the C and mC ROs (compared to the thin rectangle; $\beta_{\text{tall}} = 37.14$, 95% CI [11.98, 62.64]; $\beta_{\text{C}} = 25.76$, 95% CI [0.66, 51.13]; $\beta_{\text{mC}} = 38.76$, 95% CI [13.63, 64.46]). Crucially, these ROs also received considerably more fixations than the thin rectangle (Table 4.1). The comparison of the other ROs to the thin rectangle reveals no credible difference in RT ($\beta_{\text{thick}} = -8.85$, 95% CI [-33.99, 16.85]; $\beta_{\text{square}} = 1.02$, 95% CI [-24.40, 26.75]; $\beta_{\text{L}} = 17.52$, 95% CI [-8.36, 43.29]; $\beta_{\text{mL}} = 20.88$, 95% CI [-4.14, 46.32]). Accordingly, the higher number of fixations close to the centers of the asymmetrical ROs could be caused by a higher processing load.

Table 4.1: Absolute and relative number of fixations (a) inside the bounding boxes of the ROs (leftmost column), (b) split by left or right landing positions (left part of table), and (b) close to the center-of-mass or center-of-object of the RO (no more than 25 pixel in x or y direction, right part of table). * = For rectangular ROs, center-of-mass and center-of-object coincide. For these ROs, the numbers are the total number of fixations to their center.

	total	left		right		center-of-mass		center-of-object	
thin	979	508	51.9%	471	48.1%	63	*	63	*
thick	1197	618	51.7%	577	48.3%	25	*	25	*
square	1404	699	49.8%	705	50.2%	14	*	14	*
tall	1562	785	50.3%	777	49.7%	13	*	13	*
C	1690	810	48.0%	879	52.0%	15	34.1%	29	65.9%
mC	1665	814	48.9%	851	51.1%	27	47.4%	30	52.6%
L	1455	876	60.2%	579	39.8%	53	86.9%	8	13.1%
mL	1383	573	41.4 %	810	58.6%	34	89.5%	4	10.5%

to the more distant center-of-object comes as no surprise. To overcome this problem and to draw on a larger data set, I analyzed whether participants overall preferred to inspect the left versus the right side of the BB of each RO. A fixation landed on the left part of a BB, if its relative x coordinate was smaller than 0.5. Correspondingly, a fixation on the right part of a BB had a relative x coordinate greater than 0.5. I removed fixations where the relative x coordinate equals 0.5 (only 3 fixations). On the left part of Table 4.1, the number of fixations on each side of the BBs of the ROs are given. Here, the rectangular ROs serve as a baseline: For these symmetrical ROs, an asymmetrical gaze pattern would suggest that participants had a general left-right bias in looking behavior. This did not seem to be the case.

Considering the left-right bias for the asymmetrical ROs suggests a bias in the direction of where the mass is located for the L and mL ROs (more fixations on the side where the vertical leg of the L-shaped ROs is located) – but not for C and mC ROs which were inspected as if they were rectangular ROs. I specified a Bayesian regression model to test this observation. This model predicts the relative x coordinate of fixations (inside the BB) as a function of the RO. The fixations to the thin rectangle served as the intercept of the model, i.e., the model compared the fixations to each RO with the fixations to the thin rectangle. The outcome of the model supports my interpretation:

For the thin rectangle (the model’s intercept), the predicted average relative x coordinate of fixation is not credibly different from 0.5 ($\beta_{\text{thin}} = 0.49$, 95% CI [0.47, 0.51]). Compared to the thin rectangle, the regression model estimates no credible differences in average horizontal fixation locations for any of the other rectangles ($\beta_{\text{thick}} = -0.01$, 95%

The gaze pattern for the C-shaped ROs cannot be distinguished from the gaze pattern for the rectangular ROs. The gaze pattern for the L-shaped ROs reflects their asymmetrical mass distribution.

CI $[-0.04, 0.01]$, $\beta_{\text{square}} = 0.01$, 95% CI $[-0.02, 0.03]$, $\beta_{\text{tall}} = 0.00$, 95% CI $[-0.02, 0.03]$). This indicates no general left-right bias in participants' looking behaviors. More interestingly and despite the asymmetrical distribution of mass, the regression model does neither estimate a credible difference in horizontal fixation locations for the C and mC ROs ($\beta_{\text{C}} = 0.01$, 95% CI $[-0.01, 0.04]$, $\beta_{\text{mC}} = 0.01$, 95% CI $[-0.01, 0.03]$). This means that participants fixated the C and mC ROs as if they were rectangular. On the other hand, the model indicates that people inspected the L and mL ROs differently compared to the C and mC ROs. For these ROs, credibly more fixations landed on the side on which the vertical leg of the RO was located (left for L, right for mL, $\beta_{\text{L}} = -0.05$, 95% CI $[-0.07, -0.02]$, $\beta_{\text{mL}} = 0.06$, 95% CI $[0.03, 0.08]$).

In summary, for the L and mL ROs, the center-of-mass side was preferably fixated (compared to the cavity side) – contrasting the rating pattern for which I could not find an influence of the asymmetrical mass distribution. For the C and mC objects, a preference for the center-of-object is reflected in the gaze patterns (mirroring the rating patterns): Despite the asymmetry, participants fixated the C and mC ROs as if they were rectangular.

4.2.3 Results: Reaction Times

Reaction times are more a side product than the main outcome of the study. I neither told participants that they had to be as quick as possible nor did I tell them that their reaction time was measured. Nevertheless, the collected reaction times are interesting as an additional measure of task difficulty. Furthermore, I replicated the established finding (e.g., Carlson & Logan, 2001, note 1) that superior prepositions (like *über*, *above*) are processed faster than inferior prepositions (like *unter*, *below*) and generalized it to German prepositions. To do so, I specified a regression model predicting reaction time from used preposition (*über*, *above* vs. *unter*, *below*). This model showed a small but credible influence of preposition: If *über* (*above*) was used, participants were quicker (mean_{*über*} = 1857.73 ms) compared to *unter* (*below*, mean_{*unter*} = 1873.34 ms; $\beta_{\text{unter}} = 17.40$, 95% CI $[4.76, 29.61]$). This replication supports the overall validity of the study.

Participants were faster to judge *über* than *unter*.

Participants took longer to judge LOs on the non-corresponding side of the grazing line (compared to LOs on the corresponding side).

In the next two analyses, I was interested in whether the placements of the LOs in rows and columns affected the reaction time. In particular, participants might have taken longer for the LOs in rows R4–R7 (located on the non-corresponding side of the grazing line or on the grazing line) or for the LOs in columns C1 & C8 (not directly located above the RO, i.e., with deviating proximal orientation). The regression model that predicts reaction time from row number confirms the first hypothesis. Compared to row R1 (mean_{R1} = 1821.72 ms), reaction times were credibly longer for rows R4–R7 (mean_{R4} = 2360.34 ms, $\beta_{\text{R4}} = 192.28$, 95% CI $[149.25, 234.71]$; mean_{R5} = 2100.58 ms, $\beta_{\text{R5}} = 57.42$, 95% CI

[16.65, 98.44]; $\text{mean}_{R6} = 1953.16$ ms, $\beta_{R6} = 48.55$, 95% CI [6.65, 89.32]; $\text{mean}_{R7} = 2325.54$ ms, $\beta_{R7} = 158.88$, 95% CI [115.89, 200.74]). No other row had credibly different reaction times, except for row R2 for which people were slightly faster ($\text{mean}_{R2} = 1800.72$ ms, $\beta_{R2} = -24.59$, 95% CI [-48.91, -0.06]).

The second hypothesis that LOs in columns C1 & C8 might take longer to process was confirmed by a regression model that predicts reaction time from column. Compared to the first column C1 ($\text{mean}_{C1} = 2260.64$ ms), participants took about the same time for LOs in column C8 ($\text{mean}_{C8} = 2228.42$ ms, $\beta_{C8} = -4.09$, 95% CI [-26.52, 18.21]) but were credibly faster for columns C2–C7 ($\text{mean}_{C2} = 1762.72$ ms, $\beta_{C2} = -220.85$, 95% CI [-247.56, -194.37]; $\text{mean}_{C3} = 1646.78$ ms, $\beta_{C3} = -242.25$, 95% CI [-269.58, -215.79]; $\text{mean}_{C4} = 1559.70$ ms, $\beta_{C4} = -283.33$, 95% CI [-309.96, -257.15]; $\text{mean}_{C5} = 1547.57$ ms, $\beta_{C5} = -286.06$, 95% CI [-313.23, -259.60]; $\text{mean}_{C6} = 1653.74$ ms, $\beta_{C6} = -261.47$, 95% CI [-288.38, -234.76]; $\text{mean}_{C7} = 1759.41$ ms, $\beta_{C7} = -202.07$, 95% CI [-228.11, -175.59]).

Taken together, participants were quicker for LOs with non-deviating proximal orientation (columns C2–C7) on the side of the grazing line that corresponds to the to-be-rated preposition (rows R1–R3 & rows R8–R10) compared to LOs with deviating proximal orientation (columns C1 & C8) or LOs on the non-corresponding side of the grazing line or on the grazing line (rows R4–R7).

4.2.4 Discussion of the Empirical Study

In summary, the presented empirical study replicated known effects on spatial language understanding (different performance for *über*, *above*, vs. *unter*, *below*; grazing line effect; proximal orientation and center-of-mass orientation effects). More precisely, the study generalized these effects from English to German. These replications/generalizations provide evidence for a successfully conducted study that seamlessly integrates with earlier research (e.g., Hayward & Tarr, 1995; Logan, 1995; Regier & Carlson, 2001).

In addition, the study revealed two new empirical effects of the geometry of the RO on spatial language acceptability ratings. The first effect is the seemingly greater influence of the center-of-object orientation compared to the center-of-mass orientation – as observed with the asymmetrical ROs. In the next chapter, I present modifications to the AVS and the $\text{rAVS}_{\text{w-comb}}$ models that integrate this finding. We will see that these models perform better on the empirical data. The second effect is that relative distance modulates the two effects of proximal orientation and center-of-mass orientation. The empirical data from the rectangular ROs suggest that lower relative distance weakens (i) the effect of proximal orientation and (ii) – for high values of proximal orientation – weakens a *reversed* effect of center-of-mass orientation.

Two novel effects of geometry on acceptability ratings: center-ob-object orientation and relative distance.

Although this confirms the general prediction by the $rAVS_{w-comb}$ model (that relative distance should modulate the effects of proximal and center-of-mass orientation), the observed mechanism is different from the specific mechanism proposed in the $rAVS_{w-comb}$ model. Nevertheless this constitutes an interesting, novel finding.

In terms of eye movements as a measure of attentional deployment during spatial relation processing, the data provide evidence supporting the location of AVS's attentional focus point F (also playing an important role in the $rAVS_{w-comb}$ model): The horizontal component of participants' fixations was close to the horizontal component of the hypothesized focus point F. Although the hypothesized vertical component also matched the empirical vertical fixation locations, it remains unclear whether the preposition or the LO location triggered the fixation locations.

The eye-gaze patterns on the asymmetrical ROs revealed interesting insights into the perceptual processing of these ROs. In line with being rated almost as if they were rectangles, participants' fixations did not reflect the asymmetrical mass distribution of the C-shaped ROs. However, for the L-shaped ROs, the eye movements somewhat reflected the asymmetry with more fixations to areas that contained more RO mass. These results contribute to findings in saccadic and perceptual localization of abstract geometric shapes (e.g., Desanghere & Marotta, 2015; Melcher & Kowler, 1999; Nuthmann & Henderson, 2010; Vishwanath & Kowler, 2003). In these studies, researchers found that the center-of-mass of asymmetrical objects seems to be a preferred saccadic end point. This is consistent with the gaze pattern for the L-shaped ROs. On the other hand, the eye movements for the less-open C-shaped ROs highlight the importance of the task on eye movements in general. Here, the discussion in Vishwanath and Kowler (2003) is particularly interesting. They speculate that reference frames and spatial pooling processes similar to the weighted vector sum in AVS-like models might be important for "programming" the saccadic end point with respect to the task (see also the discussion in Melcher & Kowler, 1999).

Implications for the Directionality of the Attentional Shift

How do the empirical results help to reach the over-arching goal of my work? That is, do the data support an attentional shift from the RO to the LO as assumed by the AVS model or do they support an shift from the LO to the RO as implemented in the $rAVS_{w-comb}$ model? In terms of qualitative predictions, the findings from the asymmetrical ROs disconfirm both models. Instead of relying on the center-of-mass orientation (as implemented in both models), people seem to rely on the center-of-object orientation. I address this issue by modifying both models in the next chapter.

Considering the relative distance test case, generally speaking, the data support $rAVS_{w-comb}$'s a priori assumption that relative distance

modulates both the proximal orientation effect and the center-of-mass orientation effect. However, the specific mechanism implemented in the $rAVS_{w-comb}$ model is disconfirmed. This mechanism was inspired by the mechanism of the AVS model to accommodate the results from experiment 7 from Regier and Carlson (2001, see Section 3.2.6). The AVS model makes use of its attentional distribution to account for distance effects: Close LOs result in narrow attentional distributions and therefore, the proximal orientation is approximated. On the other hand, distant LOs evoke a broad attentional distribution that in turn approximates the center-of-mass orientation. Even though the AVS model does not explicitly mention *relative* distance, this mechanism fails to qualitatively accommodate the complex relationship of the predictors relative distance, center-of-mass orientation, and proximal orientation as observable in the empirical data.

Taken together, this brief qualitative analysis does not seem to prefer any of the two contrasting implementations of the directionality of the attentional shift – i.e., either the AVS or the $rAVS_{w-comb}$ model – over the other. In order to quantitatively assess the two models in more detail, I conducted several model simulations using the collected data and the stimuli from the study presented in this chapter. The next chapter presents these simulations with the aim to distinguish the two models in terms of their ability to accommodate the empirical results.

MODEL SIMULATIONS

The main goal of this chapter is to provide computational evidence for or against one of the two contrasting implementations of the direction of the attentional shift during spatial language verification (from RO to LO, AVS, or from LO to RO, $rAVS_{w-comb}$). To this end, this chapter reports the outcomes of several model comparison techniques applied to the cognitive models using the data and stimuli from the study presented in Chapter 4. Furthermore, this chapter introduces and assesses of two further model modifications that implement the surprising finding of seemingly greater importance of center-of-object orientation than center-of-mass orientation (see Section 4.2.1). In order to later on compare these two new center-of-object models to their center-of-mass predecessors, the chapter starts in Section 5.1 with introducing the two new modifications: the AVS bounding box (AVS-BB) model and the $rAVS$ center-of-object ($rAVS-CoO$) model.

Subsequently, Sections 5.2–5.5 present the methodology and results of four different model assessment techniques (GOF/SHO, PSP, MFA, landscaping). In Section 5.2, I tested whether the models quantitatively account for the empirical data presented in Chapter 4. To do so, I computed goodness-of-fit (GOF) and simple hold-out (SHO) values (cf. evaluation of $rAVS$ variations in Chapter 3). Following these “local” model analyses, the subsequent sections present the results from “global” model analyses (taxonomy from Pitt et al., 2006).

A “local” model analysis assesses model performance given an empirical data set. In contrast, a “global” model analysis considers the full range of model parameters – not considering empirical data. In Section 4.1.3, I already presented a “global” model analysis: the PSP method for the AVS and the $rAVS_{w-comb}$ models. Remember that this method takes stimuli as input and computes all possible model predictions. Such information is valuable as it provides details on how to falsify a model (e.g., because the model predicted a specific data pattern but later collected empirical data show a pattern that conflicts with the model prediction). Section 5.3 presents another PSP analysis for the two new model modifications introduced in Section 5.1.

The PSP method provides a qualitative measure of model flexibility: A more flexible model is able to compute more distinct patterns than a

Section 5.2:
GOF/SHO
Section 5.3: PSP
Section 5.4: MFA
Section 5.5:
landscaping

* Parts of the work presented in Chapter 5 were published in Kluth, Burigo, Schultheis, and Knoeferle (2016a, landscaping, MFA), Kluth, Burigo, Schultheis, and Knoeferle (2016b, center-of-object models, GOF and SHO for asymmetrical ROs data), Kluth et al. (2019, center-of-object models, GOF, SHO, MFA, landscaping), and Kluth and Schultheis (2018, rating distributions, Bayesian inference). This text extends on the already published details and presents a comprehensive overview of all analyses.

less flexible model. The ‘Model Flexibility Analysis’ (MFA, Veksler et al., 2015) is another “global” model analysis that provides a *quantitative* measure of model flexibility – how diverse model outcomes can be – given a set of stimuli. Section 5.4 presents an MFA based on the stimuli from the empirical study discussed in Chapter 4.

Finally, I conducted ‘landscaping’ (Navarro, Myung, Pitt, & Kim, 2003; Navarro et al., 2004) as another “global” model analysis. Landscaping asks whether a set of stimuli is informative enough to distinguish two models – i.e., whether the models in principle generate distinguishable data for the stimuli – by contrasting their fits to self- and other-model-generated data. In Section 5.5, the landscaping comparisons of several model-stimuli pairs are presented – in particular comparing the implementations of the two contrasting directionalities of the attentional shift.

After summarizing and discussing all model simulations in Section 5.6, the chapter closes with proposing a model extension in Section 5.7. This extension enables models to simulate full rating distributions instead of mean ratings and might prove useful for future research.

5.1 IMPLEMENTING THE PREFERENCE FOR THE CENTER-OF-OBJECT

This section introduces modifications to the AVS and $rAVS_{w-comb}$ models that implement the finding that people seem to prefer the center-of-object orientation over the center-of-mass orientation for spatial language verification (see Section 4.2.1). Note that the empirical study disconfirmed the specific details of the relative distance mechanism implemented in the $rAVS_{w-comb}$ model as well. The AVS model also cannot accommodate for the qualitative interactions in the empirical data. However, relative distance affects the predictors center-of-mass orientation and proximal orientation which are both central to the AVS and the $rAVS_{w-comb}$ model. Accordingly, core-parts of the models would need to be considerably changed in order to accommodate the relative distance effect. I did not pursue this path but hope that future modeling research will accommodate the effect of relative distance. For now, let us focus on implementing the center-of-object orientation.

5.1.1 *The AVS-BB Model*

The AVS model computes the center-of-mass orientation with its maximal attentional width, i.e., when all points on the RO receive the same amount of attention (see Appendix A in Regier & Carlson, 2001). I propose the following modification to the AVS model: the AVS-BB (AVS bounding box) model.¹ Instead of just considering all points of the RO, the AVS-BB model computes its vector sum using all points of the

¹ I thank Holger Schultheis for the first idea of the AVS-BB model.

bounding box (BB) of the RO. The BB of an RO is the smallest rectangle containing all points of the RO. For the rectangular ROs, the BB coincides with the RO. For the asymmetrical ROs, the BB also includes the cavities of the objects (see dashed rectangles in Figures 4.11 and 4.12, pages 94 and 95). With a uniform attentional distribution, the AVS-BB model computes the center-of-object orientation. Apart from that change, the AVS-BB model stays exactly the same as the AVS model.

This specification of the AVS-BB model might be problematic given available evidence from asymmetrical ROs with non-flat tops/bottoms facing the LO. This is because the AVS-BB model treats any asymmetric RO exactly as if it was rectangular (because the computation is based on the rectangular BB). However, this might not hold true for human processing. For instance, using an upright L-shaped asymmetrical RO for which the top that faces the LO is not flat (see Figure 3.16a on page 63), Regier and Carlson (2001, exp. 5) collected *above* rating data suggesting that people process asymmetrical ROs differently than rectangular ROs (however, the focus of the fifth experiment of Regier & Carlson, 2001, was to investigate the effect of the grazing line and not asymmetrical ROs).

It is very likely that the location of the cavity in asymmetrical ROs (i.e., whether the RO faces the LO with a flat top/bottom or not) influences spatial language evaluation in non-trivial ways. One could think of introducing another free model parameter that additionally weights the importance of vectors depending on their location inside the BB (inside or outside the RO). During the development of the predictions, however, the vector sum already showed its considerable flexibility which was confirmed by the PSP analysis. Accordingly, I will not introduce another free parameter here and leave the contrasting of different types of asymmetrical ROs for future work.

5.1.2 *The rAVS-CoO Model*

The $rAVS_{w-comb}$ model explicitly refers to the center-of-mass C in its computation (see Equation 3.6 on page 38). The implementation of a preference for the center-of-object orientation is straightforward: Instead of considering the center-of-mass C , the modified model uses the center-of-object CoO (as defined in Equation 4.1 on page 96) for its computation. I label this model the $rAVS-CoO$ ($rAVS$ center-of-object) model.

I note that the $rAVS-CoO$ model has the same issue as the AVS-BB model with asymmetrical ROs where the top/bottom facing the LO is not flat (e.g., the upright L-shaped RO from Regier & Carlson, 2001, exp. 5). Similar to the AVS-BB model, this is because the $rAVS-CoO$ model treats any asymmetrical RO as being rectangular. This might be a problematic assumption. However, for the current purpose, both the AVS-BB and the $rAVS-CoO$ model are well suited because the data

in this project was collected with asymmetrical ROs for which the LO-facing top/bottom was flat. These rating data suggest that for such stimuli people seem to ignore the cavity of the RO. Future research should investigate the influence of flat vs. non-flat tops/bottoms of ROs facing the LO more closely.

5.2 FITTING MODELS TO DATA: GOF AND SHO

As a first test of model performance, I computed how well the models accommodate the collected empirical results. To this end, I computed the GOF and SHO values for the AVS model and the $rAVS_{w-comb}$ model. In addition, I also computed these values for the two newly proposed models (AVS-BB and $rAVS-CoO$). I applied the same method as before (see Section 3.2.1), except for a small change in the lower bounds of the parameters λ and α :

$$0.001 \leq \lambda \leq 5 \quad (5.1)$$

$$0.001 \leq \alpha \leq 5 \quad (5.2)$$

Figure 5.1 presents the GOF and SHO values for different subsets of the empirical data. For Figure 5.1a, I fitted the models to the whole data set from Chapter 4. I used only half of this data set for Figures 5.1b (only ratings for LOs around rectangular ROs) and 5.1c (only ratings for LOs around the asymmetrical ROs). This subsetting provides the opportunity to separately assess the models on data from the two different test cases: relative distance and asymmetrical ROs. As a further test of the new modifications, I fitted the AVS-BB and the $rAVS-CoO$ models to the whole data set from Regier and Carlson (2001). Figure 5.1d plots these fits alongside the already known fits of the AVS and the $rAVS_{w-comb}$ model.

The performance of the two new model modifications on the data from Regier and Carlson (2001) establish their overall validity. Despite their known issue to qualitatively accommodate the data from Regier and Carlson (2001, exp. 5), the two new models perform virtually equivalent compared to the AVS and the $rAVS_{w-comb}$ model. Considering only data from the asymmetrical ROs (Figure 5.1c), however, the two new models clearly outperform their predecessors – both in GOF and in SHO. This is further evidence supporting the hypothesis that people rely more on the center-of-object orientation than on the center-of-mass orientation. Importantly, the $rAVS-CoO$ model cannot be distinguished in terms of SHO values from the AVS-BB model for the data subset from the asymmetrical ROs.

For data from the rectangular ROs (relative distance test case, Figure 5.1b), the AVS model fits the data considerably better than the $rAVS_{w-comb}$ model (GOF and SHO). Because center-of-mass and center-of-object coincide for these ROs, the AVS-BB model acts exactly like the AVS model and the $rAVS-CoO$ model acts exactly like the $rAVS_{w-comb}$

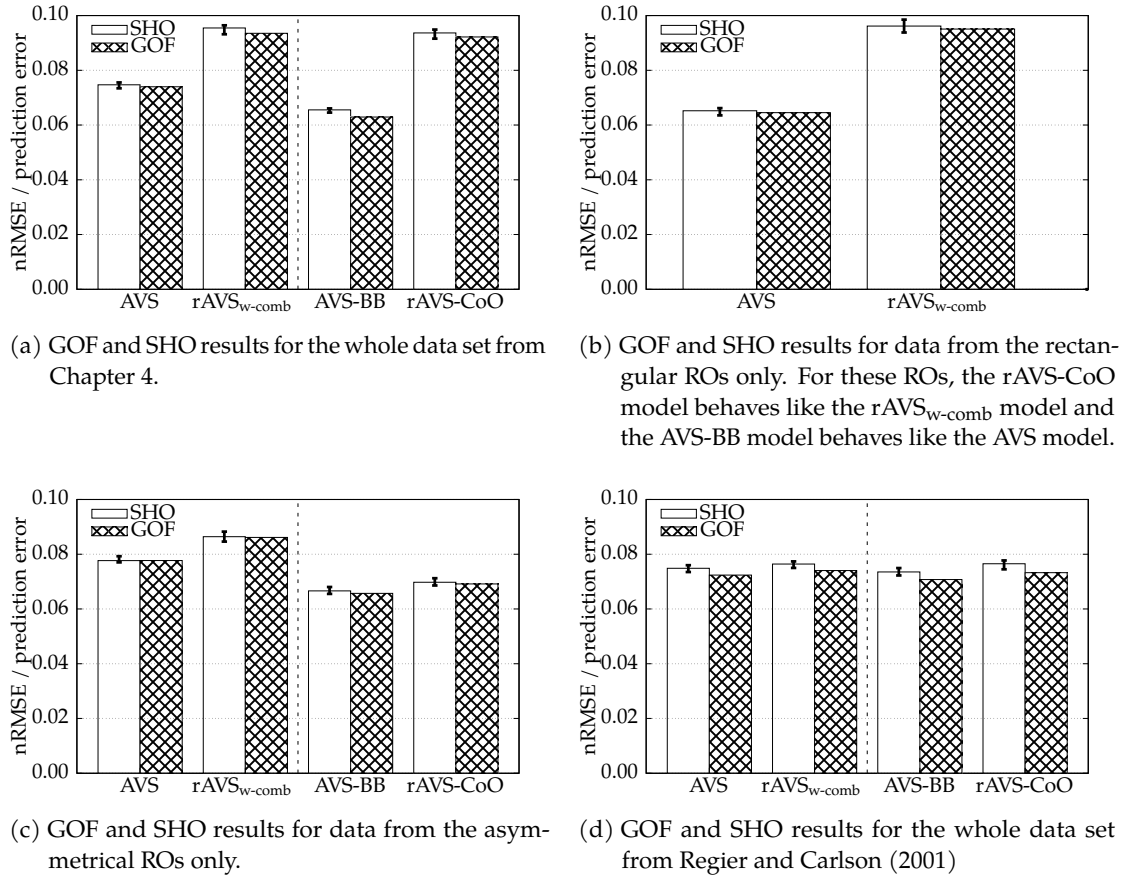


Figure 5.1: Goodness-of-fit (GOF) and simple hold-out (SHO) results for (a)–(c) the data from the study presented in Chapter 4 (collapsing across *über*, *above*, and *unter*, *below*) and (d) data from Regier and Carlson (2001). Error bars show bootstrapped 95% confidence intervals of the SHO medians. Image copyright: See Appendix E.

model. The better performance of the AVS model comes as no surprise considering that the specific relative distance mechanism as implemented in the rAVS_{w-comb} model was not confirmed by the empirical data (see Section 4.2.1). However, note that despite this quantitatively better fit, the AVS model does not qualitatively spell out how the predictors center-of-mass orientation, proximal orientation, and relative distance should interact with each other. Rather, the mechanism of the AVS model suggests a similar interaction compared with the disconfirmed mechanism in the rAVS_{w-comb} model: a close LO leads to a smaller attentional width which in turn favors the proximal orientation over the center-of-mass orientation.

The disconfirmation of the relative distance mechanism from the rAVS_{w-comb} model is also likely the reason for its worse performance (compared to the AVS model) for the whole data set (Figure 5.1a). The rAVS-CoO model inherits the relative distance mechanism from the

rAVS_{w-comb} model, explaining why it is also performing considerably worse than the AVS-BB model for the whole data set.

*The center-of-object
models outperform
the center-of-mass
models.*

Taken together, all models fit the data closely (all GOFs are below 0.10). The new models AVS-BB and rAVS-CoO outperform their predecessors for the asymmetrical ROs. The disconfirmed relative distance mechanism presumably causes the rAVS_{w-comb} and the rAVS-CoO models to perform worse than the AVS and the AVS-BB models for data from the rectangular ROs and all ROs.

5.2.1 *Motivation for Global Model Analyses*

GOF and SHO values assess model performance given a particular data set. While this is a valuable and important measurement to judge the quality of a model, it is not sufficient for a thorough model evaluation (e.g., Roberts & Pashler, 2000). Regardless of an empirical data set, it is of interest what a model can and what it cannot compute, as this gives information about how the models constrain future empirical data. Even more, “[w]ithout knowing how much a theory [model] constrains possible outcomes, you cannot know how impressed to be when observation and theory [model] are consistent” (Roberts & Pashler, 2000, p. 359). In particular, the constraints of a model (what it cannot compute) are informative for the falsification of the model.

A model that computes a wide range of data patterns (a highly flexible model) is hard to falsify because it can account for a wide range of future empirical data. Crucially, a model that performs well on a given data set (as revealed by good GOF and SHO results) might still generate such wide range of data patterns. This is because by design both the GOF and the SHO method try to restrict the ranges of the model parameters: they estimate the values of the parameters in order to provide a close fit to particular empirical data. While the GOF stops here, the SHO uses such a parameter set to compare the model output with data that are new to the model and reports this difference as result. This provides a measure of one important part of model flexibility (model generalizability; how good are the estimated parameter values for a “new” data set?) but it does not measure another important part of model flexibility: can the model generate data patterns different from the empirically observed patterns?

This is because the SHO method is a valuable “local” model analysis (how does a model perform on given data) but it cannot be used for a “global” model analysis that considers the full range of model parameters (how does the model perform in general; taxonomy from Pitt et al., 2006). For the AVS and the rAVS_{w-comb} model, I already applied a “global” model analysis to see what data patterns the models can generate: the PSP method (see Section 4.1.3). Given that the AVS-BB and the rAVS-CoO model better account for the empirical data than

the AVS and the $rAVS_{w-comb}$ model, I was interested to analyze their possible output using the PSP algorithm.

After presenting these PSP results in Section 5.3, I report the results of two further global model analyses: MFA and landscaping. In Section 5.4, the MFA helps to quantitatively investigate the flexibility of the cognitive models, answering questions such as to what extent the models generate data patterns different from the empirically observed patterns. Ideally, a model should be as flexible as needed to accommodate empirical data – but not more. Section 5.4 reports whether any of the two implemented directionalities of the attentional shift is superior in terms of model flexibility. To preview the outcomes of the MFA, the two implemented directionalities of attention cannot be reliably distinguished from each other with this additional information. This is why I conducted landscaping as the final “global” model analysis. Given a set of stimuli, this method asks whether two competing models generate data that allow a modeler to distinguish the models. The results from the landscaping method are presented in Section 5.5.

5.3 PARAMETER SPACE PARTITIONING: CENTER-OF-OBJECT MODELS

METHOD Compared to the first PSP analysis in Section 4.1.3, I used different input and a slightly different coding for this PSP analysis. This was done to better reflect the two test cases (relative distance and asymmetrical ROs; by making two instead of three rating comparisons). Furthermore, this PSP analysis draws on a greater set of LO placements compared to the first one.

More specifically, I made two comparisons (corresponding to the relative distance test case and the asymmetrical ROs test case): First, I contrasted the mean *über (above)* rating for the 28 LOs above the thin rectangle against the mean *über (above)* rating for the 28 LOs above the tall rectangle (see Figure 4.14 on page 99). A lower mean rating for LOs above the thin rectangle (vs. tall rectangle) is coded as “-”, a higher mean rating for LOs above the thin rectangle (vs. tall rectangle) is coded as “+”. Equal mean ratings are coded as “0”. Second, I contrasted two mean *über (above)* ratings for 12 LOs above the L-shaped RO: 6 LOs to the left of the center-of-mass of the RO (columns C2–C3, Figure 4.12, page 95) against 6 LOs to the right of the center-of-mass of the RO (columns C4–C5). Here, a “-” codes for a lower mean rating for the left LO-set compared to the right LO-set and a “+” codes for a higher mean rating for the left LO-set compared to the right LO-set. A “0” denotes no difference in mean ratings. To define equality of mean ratings, I used the two equality thresholds $t_e \in \{0.1, 0.5\}$. The full PSP pattern is thus a two-digit code: The first digit codes the difference in mean ratings for LOs above the thin vs. tall rectangle, the second digit codes the difference in mean ratings for LOs to the left of the center-of-mass vs. to the right of the center-of-mass of the L RO.

“+”: *first > second*
 “-”: *first < second*
 “0”: *first = second*

RESULTS Figure 5.2 plots the mean relative volume estimates from three PSP runs for the two different thresholds (Figure 5.2a: $t_e = 0.1$; Figure 5.2b: $t_e = 0.5$). Due to the slightly different pattern coding and to ease comparison, I computed this PSP analysis for both: the two new models and their predecessors.

Across a substantial number of parameter settings, the center-of-object models generate the empirical pattern for the asymmetrical ROs.

Considering the motivation of the two new models (implementing a preference for the center-of-object orientation), their development was successful. This can be seen from the second PSP digit (coding for LOs above the asymmetrical L RO). Here, the empirical pattern is a “-” because participants rated LOs more central with respect to the center-of-object (the right LO-set in the PSP input) considerably higher than less central LOs (the left LO-set). The two new models confirm this pattern: For both thresholds t_e and for both new models, the vast majority of the parameter space is covered by patterns where the second digit is a “-” (i.e., patterns “-” and “0-”).

In contrast, the AVS and the $rAVS_{w-comb}$ models cannot accommodate the empirical pattern from LOs around the asymmetrical ROs. Given that the center-of-object orientation effect was relatively large (see Section 4.2.1), the threshold $t_e = 0.1$ appears to be too small.² However, for the threshold $t_e = 0.5$ (Figure 5.2b), only the AVS model computes the empirical pattern for the asymmetrical ROs – albeit, with a small volume (< 4%), clearly not being a central outcome of the AVS model.

Neither model fully accommodates the relative distance test case. Future modeling research should address this more closely.

For the relative distance test case, the empirical pattern is “0”: no difference in ratings across rectangles with different heights. While all models generate this sub-pattern (i.e., patterns “0-” and “00”) to some extent for all thresholds t_e , only the AVS and the AVS-BB models generate it with $t_e = 0.5$ in a majority in their parameter space (> 50%, Figure 5.2b). However, the complete empirical pattern is “0-”: no difference in ratings across rectangles *and* lower ratings for LOs less central with respect to the center-of-object. This pattern is only generated by the AVS and the AVS-BB model. For the AVS model it is obviously not a main prediction: For $t_e = 0.1$ (Figure 5.2a), it occupies less than 4% of its parameter space, and for $t_e = 0.5$ (Figure 5.2b), it is virtually non-existent (< 0.3%). The AVS-BB model generates the empirical pattern with a greater set of parameters. Nevertheless, the volumes covered in its parameter space are still comparably small. For $t_e = 0.1$, the AVS-BB model generates the pattern “0-” for < 5% of its parameter space, and for $t_e = 0.5$, the volume is < 16%.

These results again demand more modeling efforts for the relative distance effect – a task that goes beyond this Ph.D. project, as central

² It is difficult to compare the outcome of ordinal regression models to differences in mean ratings. This is because ordinal data should not be treated as metric data. However, since all cognitive models considered in this thesis make this “mistake”, I kept it for the PSP analysis. Future research should run PSP analyses with extended models that consider ratings as ordinal data. See Section 5.7 for first steps in this direction.

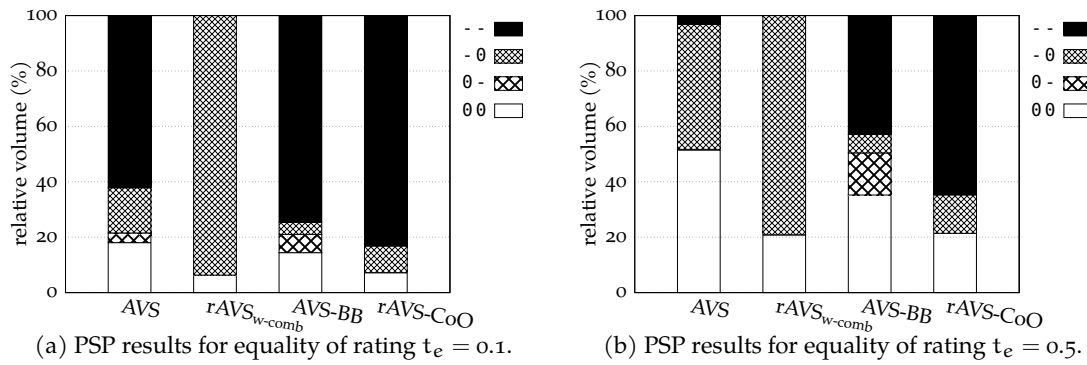


Figure 5.2: Results of the second PSP analysis: Estimations of relative volumes in parameter spaces of the models covered by distinct qualitative patterns (averaged over three PSP runs). First digit codes for difference in mean *über* (*above*) ratings for 28 LOs above the thin rectangle vs. the tall rectangle. Second digit codes for difference in mean *über* (*above*) ratings for 6 LOs to the left vs. to the right of the center-of-mass of the L-shaped RO. Mean ratings were considered equal if they differed less than (a) $t_e = 0.1$ or (b) $t_e = 0.5$. Image copyright: See Appendix E.

parts of the models have to be changed to properly address the effect. Within this Ph.D. project, however, I thoroughly analyzed the present models further. In the next section, I report the results of the recently proposed MFA – another global model analysis.

5.4 MODEL FLEXIBILITY ANALYSIS

The ‘Model Flexibility Analysis’ (MFA, Veksler et al., 2015 provides a quantitative measure of the flexibility of a model. The flexibility of a model is defined as its ability to produce arbitrary data. The more flexible a model is, the greater is its range of possible output. A model needs to have some flexibility but it should not be too flexible. High flexibility may result in (i) predicting more qualitatively different patterns (see PSP analyses), (ii) over-fitting empirical data (see SHO results), or (iii) mimicking other models (see forthcoming landscaping analysis, Section 5.5).

These problems can be (at least) measured by applying the methods just mentioned in parentheses. However, each method accounts for a single potential problem of model flexibility only. Moreover, each method considers model flexibility only indirectly by measuring symptoms. For instance, a more flexible model tends to over-fit data, so SHO checks for over-fitting instead of measuring model flexibility. One reason for this indirect approach to model flexibility is the fact that several measures of model flexibility exist (see Veksler et al., 2015, for an overview including relations to the MFA).

A model needs to have some flexibility but it should not be too flexible.

I chose the MFA as it directly relates the actual model output to the *theoretically* possible output, i.e., the MFA actually implements the very definition of model flexibility in a straightforward way. However, this comes with a cost. To do an MFA, one must enumerate all possible model predictions which is often virtually impossible or at least very time-consuming (one reason why the PSP analysis was developed is the huge amount of computations needed to enumerate the whole parameter space).

5.4.1 Model Flexibility Analysis: Method

The outcome of the MFA is the value ϕ which denotes the model flexibility on a scale from 0 (low flexibility) to 1 (high flexibility). More precisely, ϕ is the number of all outputs a model is able to generate divided by all theoretically possible patterns:

$$\phi = \frac{\text{number of data patterns the model can generate}}{\text{number of all theoretically possible data patterns}} \quad (5.3)$$

A high value of ϕ means that the model is highly flexible (it generates almost all theoretically possible data). A low value of ϕ means the opposite: the model has a low flexibility (it generates only a small subset of all possible data).

To make things concrete, consider an RO with two LOs placed above it. The number of theoretically possible data patterns is $9^2 = 81$ (range of rating scale, 9, to the power of the dimension of one data pattern, 2 ratings). That is, there are 81 possible data patterns for this simple example: (1,1), (1,2), (1,3), etc. For almost all tasks, the space of all possible data is larger than the space of empirically plausible data. In our example, the two LOs are placed, say, very central above the RO. Human rating data then would mostly consist of high ratings, say, 7 or higher (7, 8, 9: 3 plausible ratings). This makes the number of empirically *plausible* data $3^2 = 9$. A model that predicts ratings between 4 and 9 (6 possible ratings) is then able to compute all plausible ratings but also non-plausible ratings (lower than 7). For this model, the number of distinct model outputs is $6^2 = 36$ and accordingly MFA's $\phi = \frac{36}{81} \approx 0.44$. This means that the model is able to generate ~44% of all possible data patterns. This model is more flexible than a model that only predicts ratings between 7 and 9 ($\phi = \frac{3^2}{81} \approx 0.11$). Crucially, both exemplary models could fit the empirical data equally well (they both generate the empirical data) – but the lower model flexibility of the second model would favor it over the first. This is because the second model does not predict empirically implausible data (ratings lower than 7).

The MFA computes the ratio ϕ by enumerating the whole parameter space of a model, (i.e., all possible parameter settings) and generating a data pattern for each parameter setting. However, both the model

parameters and the data patterns are continuous variables.³ This is why I need to specify two granularities: one for the parameter space (denoted with j) and one for the data space. All models discussed here have four free parameters. I split every parameter range into $j = 50$ intervals which leads to a total of 50^4 data patterns generated by each model. Each of these data patterns consists of “mean” ratings for all LO locations used as input. Since I considered different subsets of the data presented in Chapter 4, these data patterns have different lengths. For the entire data set, the number of mean ratings (i.e., the dimension) of every data pattern is $28 \text{ LOs} \times 8 \text{ ROs} \times 2 \text{ prepositions} = 448$.

After generating 50^4 data patterns (each with the length 448), the MFA compares how similar the data patterns are. In other words, the MFA determines the area of the data space covered by these data patterns. To this end, the second granularity comes into play: a “grid” that splits the data space into cells. Here, Veksler et al. (2015) suggest to use $\sqrt[n]{j^k}$ cells for each dimension in the data space (n is the dimension of one data pattern; j is the granularity of the parameter space; k is the number of parameters; for my case: $\sqrt[48]{50^4}$ cells). All model-generated data patterns that fall into the same cells across all data dimensions are considered equal, i.e., they are counted only once. Data patterns that are falling into different cells in the data space are considered unequal. The number of such unequal data patterns is the total number of distinct data patterns a model generates. This number (the numerator of Equation 5.3) is finally divided by the number of all cells in the data-space-grid (the denominator of Equation 5.3). Using the $\sqrt[n]{j^k}$ grid for all n dimensions as suggested by Veksler et al. (2015), the total number of grid cells equals the total number of generated model outcomes: $\left(\sqrt[n]{j^k}\right)^n = j^k$.

If every parameter setting produces a distinct model output (i.e., each model output falls into a different cell than all other outputs), the number of unique model outputs is equal to the number of all cells resulting in $\phi = \frac{50^4}{50^4} = 1$. Such a model is maximally flexible. The more model outputs are considered equal, the smaller becomes the area covered by the model outcomes. The lowest possible value⁴ for ϕ for the whole data set would be $\phi = \frac{1}{50^4} = 1.6 \times 10^{-7}$. This lowest possible value would be produced by a model that generates the same output regardless of its parameter setting. Such a model is maximally inflexible.

3 For the data patterns the same caveat as before is valid: Rating data should be treated as ordinal (i.e., discrete) data, not as metric (i.e., continuous) data. However, all considered models compute continuous “mean” ratings. This is why I also had to conduct the MFA with continuous data patterns.

4 There can be no value $\phi = 0$ because this would mean that the model produces no output at all.

The Curse of Dimensionality

After having introduced the MFA as proposed by Veksler et al. (2015), I note problems with the suggestion to split every dimension of the data space into $\sqrt[4]{50^4}$ cells. Applied to the data from Chapter 4 with $n = 448$ dimensions, this results in ca. 1.04 cells per dimension. Every dimension consists of a mean rating (for a single LO) on a rating scale from 1 to 9. Splitting this rating scale into 1.04 cells means that two unequally sized intervals emerge: The first interval ranges from rating 1 to rating 8.64, the second interval ranges from 8.64 to 9. Crucially, only mean ratings that fall into different intervals are considered to be different. For the data from Chapter 4 this means that all mean ratings between 1 and 8.64 are treated as being equal. This is obviously a problematic assumption.

There are two possibilities to enlarge the number of intervals per dimension in the data space while still keeping the suggestion of $\sqrt[4]{j^4}$ cells per dimension: First, increase j (i.e., compute more model outputs by using a finer granularity for every model parameter) and second, decrease n (i.e., choose a smaller data set). Unfortunately, the first option is infeasible. To obtain a reasonable number of intervals in every dimension of the data space, say 3, I would need to split every parameter range into $j = \sqrt[4]{3^{448}} \approx 2.74 \times 10^{53}$ intervals, which is magnitudes greater than what is currently possible. The computation with $j = 50$ already takes several days and consumes large amounts of computing resources (and I would like to compute several models on several subsets). An attempt to tackle the problem by choosing the second option – reducing the number of dimensions in the data set by aggregating data – failed, too.⁵

There is a third possibility to enlarge the number of cells in the data space without aggregating data, though. This third possibility, however, does not follow the $\sqrt[4]{j^4}$ suggestion by Veksler et al. (2015). In contrast to this rather arbitrary partition of the data space in terms of the domain-specific meaning of the data space dimensions⁶, the third possibility applies a partition that is sensible with respect to the domain of the data. More specifically, it uses a domain-specific value as the number of cells per dimension: the range of the rating scale. This provides a

⁵ I did this by aggregating model predictions as mean ratings for distinct ROs (i.e., only one mean rating per RO, collapsing across several mean ratings for single LOs). I have $n = 10$ ROs (four rectangles, two C-shaped objects and four L-shaped objects) which results then in $\sqrt[4]{50^4} \approx 4.78$ cells per dimension of the data space. Running the MFA on these summarized data (one mean rating per RO), however, resulted in different relative rankings of model flexibility compared to using mean ratings for single LOs placed around the ROs. Possibly, this is due to the caveats of the MFA mentioned by Evans, Howard, Heathcote, and Brown (2017) who reported that using different data summary statistics changes the results of the MFA. Since the models were developed to account for the mean rating of a single LO and not the mean rating collapsing across several LOs, I neither report nor discuss the MFA on this condensed data set further.

⁶ The suggestion from Veksler et al. (2015) makes sense with respect to obtaining an easily interpretable ϕ value.

more fine-grained partition of the data compared to the problematic suggestion from Veksler et al. (2015): Instead of treating ratings from 1 to 8.46 as equal (see above), the domain-specific partition distinguishes all 9 ratings from each other.

However, this approach has the side effect that the data space is split in more cells (9^{448}) than the number of generated model outcomes (50^4). This means that $\phi = 1.0$ can never happen: Even if all model outputs are different from each other, the maximum value for ϕ is $\phi_{\max} = \frac{50^4}{9^{448}} \approx \frac{6.25 \times 10^6}{3.17 \times 10^{427}} = \frac{1}{3.17 \times 10^{421}} = 3.17 \times 10^{-421}$. This makes interpreting the absolute value of ϕ difficult, because its possible values are not ranging from 0.0 to 1.0 anymore. In order to account for this, I report below all maximal possible values ϕ_{\max} of ϕ as well as the normalized $\phi_n = \frac{\phi}{\phi_{\max}}$.

Taken together, I report three different MFA ratios: ϕ_1 , ϕ_2 , and ϕ_{n2} . ϕ_1 was computed with $\sqrt[4]{50^4}$ cells in every dimension of the data space. For ϕ_2 , I chose the range of the rating scale as number of cells per dimension of the data space (i.e., 9 cells for the data from Chapter 4). $\phi_{n2} = \frac{\phi_2}{\phi_{2\max}}$ is the normalized version of ϕ_2 . The flexibility of a model is always defined with respect to the stimuli used as input. I used four different stimuli sets to compute the MFA ϕ s: the whole stimuli set from Chapter 4, the rectangular ROs only, the asymmetrical ROs only, and the whole stimuli set from Regier and Carlson (2001).

I used the same parameter ranges as for the other simulation methods, see Equations 3.15–3.17 (page 42) and 5.1–5.2 (page 118). Evans et al. (2017) recently criticized the MFA for producing invariant model flexibilities with different parameter ranges. Therefore, I also re-computed all MFA results using smaller but still plausible parameter ranges for some parameters:

$$0.0005 \leq \lambda \leq 3.0 \quad (5.4)$$

$$0.0005 \leq \alpha \leq 3.0 \quad (5.5)$$

$$0.0 \leq \text{highgain} \leq 2.0 \quad (5.6)$$

Surprisingly, I found higher flexibilities for these smaller parameter ranges than for the larger parameter ranges (see below). This confirms parts of the critique by Evans et al. (2017). Accordingly, I do not discuss the *absolute* values of the ϕ values. However, given that the relative rankings of the computed flexibilities did not change (with one exception that I discuss below), I still think that at least MFA's relative flexibilities are an interesting and valuable measure of model flexibility. Since Evans et al. (2017) report other problems of the MFA, the MFA results should be interpreted with caution and related to the outcomes of other methods measuring model flexibility (e.g., PSP or landscaping).

5.4.2 Model Flexibility Analysis: Results

Tables C.1 and C.2 in the appendix list all MFA ϕ values. Figure 5.3 plots most of these numbers. I plotted the ϕ_1 and ϕ_{n_2} values, i.e., the results computed with the number of cells per data space dimensions as suggested by Veksler et al. (2015, ϕ_1 , Figure 5.3a) and the results where the number of cells per data space dimension equals the range of the rating scale, normalized with the respective $\phi_{2\max}$ (ϕ_{n_2} , Figure 5.3b; see Tables C.1 and C.2 for the $\phi_{2\max}$ values). In addition, I also plotted the corresponding MFA results that I computed with smaller parameter ranges (Equations 5.4–5.6) as a double-check for the critique by Evans et al. (2017, Figures 5.3c and 5.3d).

Validity of MFA Results

Before I discuss the MFA results in terms of model flexibility, I relate ϕ_{n_2} to ϕ_1 in order to test the validity of my approach circumventing the problem of sparse data space cells for ϕ_1 . Moreover, I address one particular critique point raised by Evans et al. (2017).

Using a domain-specific data space grid and normalizing ϕ leads to comparable relative flexibility estimates.

Comparing ϕ_1 with ϕ_{n_2} (i.e., Figure 5.3a vs. 5.3b) yields the same relative flexibilities for all models and stimuli sets.⁷ There is only one exception: While $rAVS_{w\text{-comb}}$'s ϕ_1 for the subset consisting of the asymmetrical ROs is lower than all ϕ_1 s for the stimuli from Regier and Carlson (2001, see Figure 5.3a), $rAVS_{w\text{-comb}}$'s ϕ_{n_2} for the asymmetrical ROs is higher than the ϕ_{n_2} s for the stimuli from Regier and Carlson (2001, see Figure 5.3b). However, the difference is only small and arises only across stimuli sets, probably rendering it not important.

Evans et al. (2017) reported that the MFA computes invariant model flexibilities with different parameter ranges. To analyze whether this is a problem for the models and stimuli investigated here, I computed the MFA results with narrower parameter ranges for the same model-stimuli pairs. These results are plotted in Figures 5.3c and 5.3d. First of all, the absolute values of ϕ_1 and ϕ_{n_2} are larger than their counterparts computed with a greater range of parameters. This is unexpected if not to say “unambiguously incorrect [...], as a wider range of parameter values allows for a greater range of predictions” (Evans et al., 2017, p. 342) and hence a smaller range should lead to *lower* flexibility estimates. For most ϕ_{n_2} , the MFA result is more than double in size for the smaller range compared to the larger range (cf. y-axes of Figures 5.3b vs. 5.3d). For ϕ_1 , the difference is not that exaggerated, though (Figures 5.3a vs. 5.3c). This calls the validity of the absolute values of ϕ into question.

⁷ If one looks at the non-normalized values of ϕ_2 (not plotted, see Tables C.1 and C.2), however, the relative flexibilities across stimuli sets are not preserved: a higher dimensionality of the data space automatically leads to lower results for ϕ_2 . This illustrates the need for normalizing the ϕ_2 value which basically relates the dimensionality of the data space to the absolute outcome of ϕ_2 .

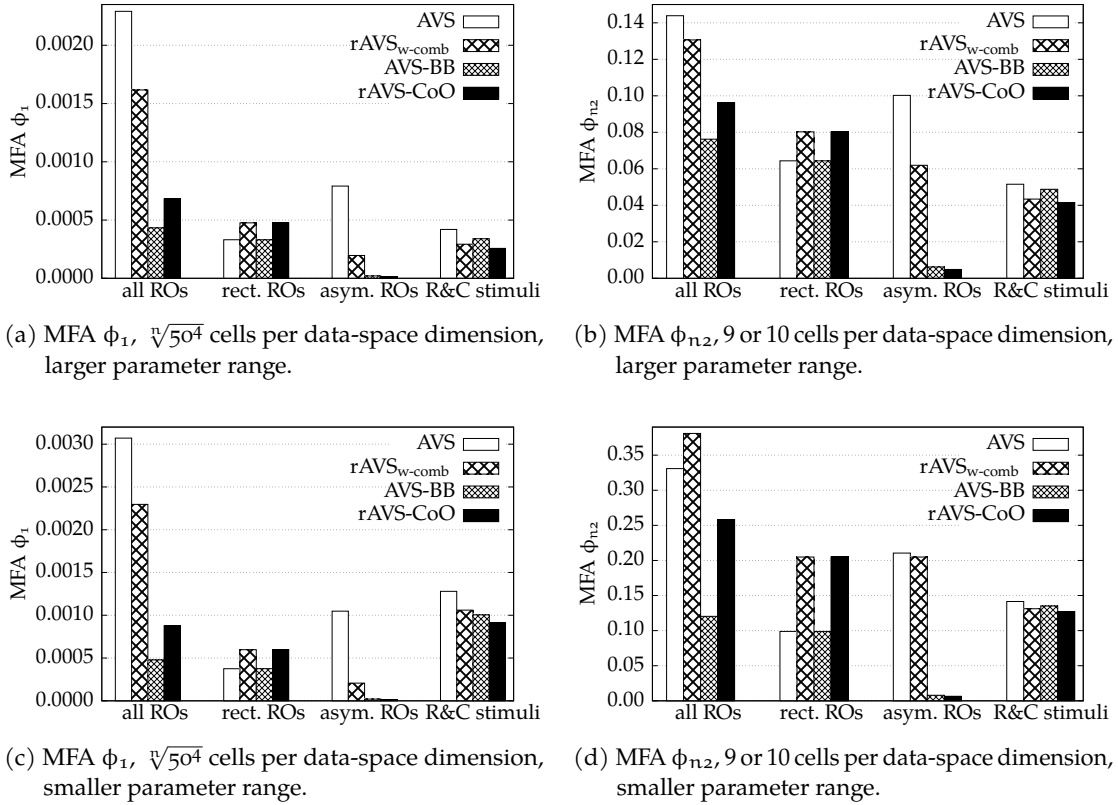


Figure 5.3: Results of the Model Flexibility Analysis (MFA). The lower ϕ , the less flexible is the model. Note the different y-axes. Panels (a) and (c) show ϕ_1 , i.e., results computed with the number of data space cells as suggested by Veksler et al. (2015). Panels (b) and (d) show ϕ_{n_2} , i.e., results computed with as many cells for every data-space dimension as there were rating intervals (i.e., 9 for the stimuli from Chapter 4, 10 for stimuli from Regier & Carlson, 2001, which are abbreviated as R&C stimuli in the plots) and normalized by dividing with the corresponding ϕ_{2max} . See Tables C.1 and C.2 for more results. For panels (c) and (d), I used smaller parameter ranges (see Equations 5.4–5.6) to address parts the MFA-critique by Evans et al. (2017).

However, in terms of relative model flexibilities, the computations with smaller parameter ranges replicate the first computations almost entirely: Although across stimuli sets the computed relative flexibilities are not always stable (e.g., rectangular ROs vs. stimuli from Regier & Carlson, 2001, in Figures 5.3a vs. 5.3c), the relative model rankings *within* stimuli sets are reproduced with a single exception only. This provides support for the validity of relative model flexibilities as computed by the MFA. The exception concerns the ϕ_{n_2} flexibilities of the AVS vs. the $rAVS_{w-comb}$ model for the whole stimuli set from Chapter 4 (cf. Figures 5.3b vs. 5.3d).

While the absolute values of ϕ should not be interpreted, the relative model flexibilities appear to be sound.

Relative Model Flexibilities

In terms of model flexibility, the AVS and the rAVS_{w-comb} model cannot be distinguished reliably.

Considering this exception (flipped relative ϕ_{n2} flexibilities for AVS and rAVS_{w-comb} for the whole stimuli set) together with (i) varying relative flexibilities of the AVS and the rAVS_{w-comb} models for other stimuli sets (rAVS_{w-comb} is more flexible than AVS for rectangular ROs but less flexible for other stimuli) and (ii) the general caution one should exercise when interpreting MFA results (Evans et al., 2017), the MFA results cannot be used to reliably distinguish the AVS and the rAVS_{w-comb} models in terms of their flexibility.

More interesting with respect to the finding that people seem to prefer the center-of-object orientation over the center-of-mass orientation is the fact that the two center-of-object models AVS-BB and rAVS-CoO are less flexible for all stimuli sets than their predecessors. This further supports the two new models: Despite their lower flexibility, they fit the empirical data equally well or even better than the AVS and the rAVS_{w-comb} models (cf. GOF and SHO results in Figure 5.1 on page 119). In particular for the asymmetrical ROs for which the center-of-object effect originated, the two center-of-object models are considerably less flexible than the center-of-mass models. For the whole stimuli set, this difference is weaker but still pronounced.

In terms of model flexibility, the AVS-BB and the rAVS-CoO model cannot be distinguished reliably.

However, unfortunately, the current MFA results do not shed much light on potential answers to my main research question (whether attention shifts from the RO to the LO, AVS-BB, or from the LO to the RO, rAVS-CoO). This is because in terms of model flexibility neither the AVS-BB nor the rAVS-CoO model performs substantially better than the other model: the rAVS-CoO model is less flexible than the AVS-BB model for the asymmetrical ROs and the stimuli from Regier and Carlson (2001); the AVS-BB model is less flexible than the rAVS-CoO model for the whole stimuli set⁸. While the corresponding benefit is greater for the AVS-BB model than for the rAVS-CoO model, the general issues with the MFA and specifically the flipped rankings do not allow to announce a “winning” model. The PSP results (see Figure 5.2 on page 123) support this conclusion as they – in contrast to lower ϕ 's for AVS-BB – slightly favor the rAVS-CoO model as less flexible (3 patterns vs. 4 patterns). Accordingly, the MFA results do not speak for or against any of the implemented conflicting assumptions about the directionality of the attentional shift.

After all, are the models distinguishable on the existing stimuli? To answer this question and to provide a final perspective on the models, I present several model comparisons using the landscaping method (Navarro et al., 2003, 2004) in the next section. Given a specific set of stimuli, landscaping reveals information about the potential to dis-

⁸ and the rectangular ROs – but for these ROs, the center-of-mass and center-of-object coincide, i.e., the AVS-BB model behaves like the AVS model and the rAVS-CoO model behaves like the rAVS_{w-comb} model.

tinguish two models by contrasting their fits to self- and other-model-generated data.

5.5 LANDSCAPING

Landscaping (Navarro et al., 2003, 2004) shows the relative performances of two models given an experimental design. If two models generate similar data given this design, one cannot distinguish the two models from each other. In terms of GOF and SHO results, the AVS model fits the whole data set better than the $rAVS_{w-comb}$ model and the AVS-BB model and the $rAVS-CoO$ model perform better on the data for the asymmetrical ROs than the unmodified models (see Figure 5.1 on page 119). Why then do I apply a tool intended to help with distinguishing two models?

I do this for two reasons: First, landscaping provides a qualitative measure of model flexibility. Together with the PSP results, this is complimentary evidence for the model flexibilities as computed with the criticized MFA. Second, I found equal performance of the following models on some data sets that disappear on other data sets: the $rAVS_{w-comb}$ model and the $rAVS-CoO$ model for the whole data set from Chapter 4, the AVS-BB model and the $rAVS-CoO$ model for the asymmetrical ROs, and all models for the data from Regier and Carlson (2001). A global model analysis such as landscaping might provide insights into the nature of these equal performances. These insights are specified as “the informativeness of a data set in deciding between [competing models]” (Navarro et al., 2004, p. 48). In particular the non-distinguishable performances of the $rAVS-CoO$ and the AVS-BB model are crucial to provide an answer for my main research question (directionality of attention): Given that the $rAVS-CoO$ and the AVS-BB model implement contrasting assumptions about the directionality of the attentional shift, investigating why they perform comparable on some data sets sheds light on the mechanisms of both shift implementations.

5.5.1 *Landscaping: Method*

What is landscaping then? A landscape consists of model fits (GOFs) to several artificial data sets generated by two models given a constant set of stimuli. For the generation of one artificial data set, model parameters are randomly sampled from a uniform distribution across the parameter ranges (Equations 3.15–3.17, page 42, and 5.1–5.2, page 118). This random parameter set is used to generate an artificial data set to which a small amount of noise is added. I used Gaussian noise with a standard deviation of 0.3, based on the magnitude of the standard error of the mean for the whole empirical data set from Chapter 4. I conducted the landscaping method with 1000 artificial data sets per model (i.e., 2000

artificial data sets per stimuli set). Each of the two models is fitted to each artificial data set – both self-generated data sets as well as the data sets generated by the other model.

Given that a model naturally fits self-generated data well, its ability of fitting data generated by a different model can be interpreted as what Wagenmakers, Ratcliff, Gomez, and Iverson (2004) call *model mimicry*: the ability of a model to mimic a different model. Consider two hypothetical models A and B. Further suppose that model A mimics model B but model B does not mimic model A. This means that model A is more flexible than model B because model A (i) fits the data generated from model B (suggesting that model A is able to generate data similar to model-B-generated data) and (ii) generates a wider range of data than model B (because model B cannot fit model-A-generated data).

Wagenmakers et al. (2004) proposed to use the ‘Parametric Bootstrap Cross-Fitting Method’ (PBCM) to measure model mimicry. The data-uninformed version of the PBCM is the same as the landscaping method with two exceptions: In the PBCM no noise is added to the artificial data sets and the results of the two methods are plotted differently (histograms of fit-differences for PBCM; model fits against each other in so-called landscape plots for landscaping). I followed the landscaping procedure⁹ (i.e., I added noise to the artificial data) but plotted the results using both types of plots, as this provides valuable additional perspectives on the outcomes of the simulations.

5.5.2 Landscaping: Results

In total, I conducted five landscaping analyses: In light of equal performance on the corresponding data sets, I contrasted the rAVS-CoO and AVS-BB models (i) on the asymmetrical ROs and (ii) the stimuli from Regier and Carlson (2001). The rAVS_{w-comb} model and the rAVS-CoO model performed comparable on the whole data set from Chapter 4 – despite the better performance of the rAVS-CoO model on the data from the asymmetrical ROs. This is why I ran two landscaping analyses with the rAVS_{w-comb} model and the rAVS-CoO model using these two stimuli sets. Finally, I computed a landscaping analysis with the rAVS and the AVS model for the stimuli from Regier and Carlson (2001).

rAVS-CoO vs. AVS-BB The results of the two landscaping analyses contrasting the rAVS-CoO and AVS-BB models are plotted in Figures 5.4 (asymmetrical ROs) and 5.5 (stimuli from Regier & Carlson, 2001). In

- rAVS-CoO vs. AVS-BB:
- asym. ROs
- R & C ROs
- rAVS_{w-comb} vs. rAVS-CoO:
- all ROs
- asym. ROs
- rAVS_{w-comb} vs. AVS:
- R & C ROs

⁹ Navarro et al. (2004) developed landscaping for statistical models, i.e., models that work with probability functions and not necessarily produce the same output with the same set of parameters. This allows to compute an index across all model fits to artificial data that gives information about the extent to which one model fits the data better than the other. However, since all models in this thesis are deterministic models (they produce the same output for the same set of parameters), it is not possible to compute this index (see Section 5.7 for an extension that makes the models probabilistic).

both Figures 5.4 and 5.5, panels (b) and (c) show the landscape plots (Navarro et al., 2004). Panels 5.4b & 5.5b show fits to data generated by the rAVS-CoO model, Panels 5.4c & 5.5c visualize fits to AVS-BB-generated data. The dashed line in the landscape plots is the line of equal fit, the asterisks depict the GOF results to the empirical data (see Figure 5.1 on page 119). The data from each landscape in Panels 5.4b, 5.4c, 5.5b, or 5.5c appear as one histogram of GOF-differences in Panels 5.4a and 5.5a – the plot type for the PBCM (Wagenmakers et al., 2004). For these histograms, all corresponding fits (i.e., two fits to the same data set) were subtracted from each other (GOF rAVS-CoO – GOF AVS-BB) and binned into a histogram (with black or white histogram bars; the legends in the plots relate the color to the identity of the data-generating model).

Inspecting the landscape plots in Panels 5.4b, 5.4c, 5.5b, and 5.5c reveals that the two models rAVS-CoO and AVS-BB are mimicking each other only to a small extent: Almost all model fits lie in the region of the data-generating model. That is, if the rAVS-CoO model generates data, it mostly fits these data better than the AVS-BB model (fits are to the upper left of the line of equal fit in Panels 5.4b & 5.5b). In contrast, if the AVS-BB model generates data, it mostly fits these data better than the rAVS-CoO model (fits are to the lower bottom of the line of equal fit in Panels 5.4c & 5.5c).

Nevertheless, both models fit the not-self-generated data well, as is evident from the overall low nRMSE values in the landscape plots. Moreover, all histograms in Panels 5.4a and 5.5a peak around 0.0. This corresponds to the higher density of points close to the line of equal fit (vs. farther away) in the landscape plots. This result indicates that for some generated data the models do mimic each other: Despite being generated by a different model, both models provide a comparable fit. However, this is only true for a subset of the generated data: The histograms are only extended into the direction that corresponds to better fits for the data-generating model (mirroring the direction of the “tails” in the landscape plots).

Comparing the landscaping analyses for the two different stimuli sets with each other reveals that for the stimuli from Regier and Carlson (2001) the models are mimicking each other even less than for the asymmetrical ROs (smaller histogram peaks around 0.0 and longer landscape tails in Figure 5.5 vs. Figure 5.4). Although the model mimicry is limited, for both stimuli sets, the AVS-BB model shows a higher degree of model mimicry compared to the rAVS-CoO model. That is, the AVS-BB model fits rAVS-CoO-generated data slightly better than the rAVS-CoO model fits AVS-BB-generated data. This is reflected in shorter histogram and landscape tails for rAVS-CoO-generated data (Panels 5.4b & 5.5b or black histograms in Panels 5.4a & 5.5a) compared to the AVS-BB-generated data (Panels 5.4c & 5.5c or white histograms in Panels 5.4a & 5.5a). Given that model mimicry is a measure of model

The rAVS-CoO and the AVS-BB model mimic each other only to a small extent. The AVS-BB model mimics the rAVS-CoO model stronger than vice versa.

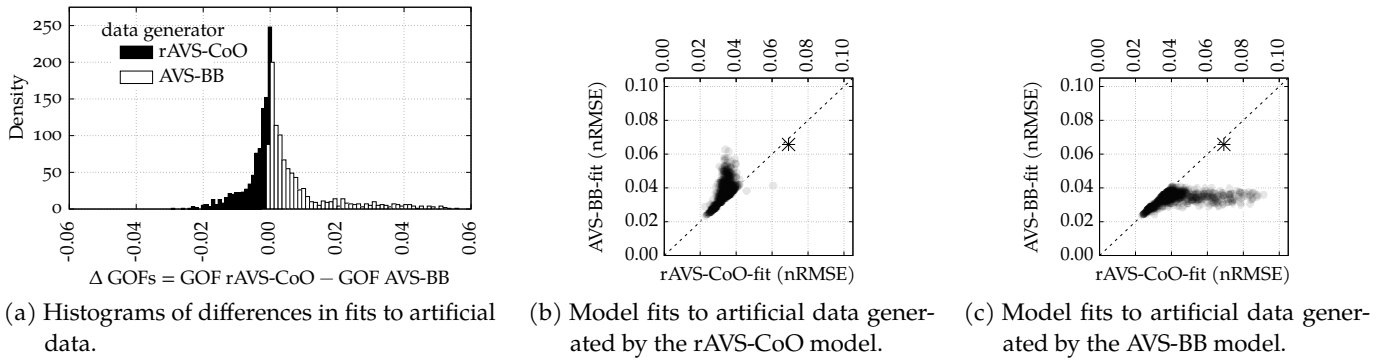


Figure 5.4: Landscaping results contrasting the rAVS-CoO model with the AVS-BB model on the asymmetrical ROs (collapsing across *über*, *above*, and *unter*, *below*). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 5.1c). Image copyright: See Appendix E.

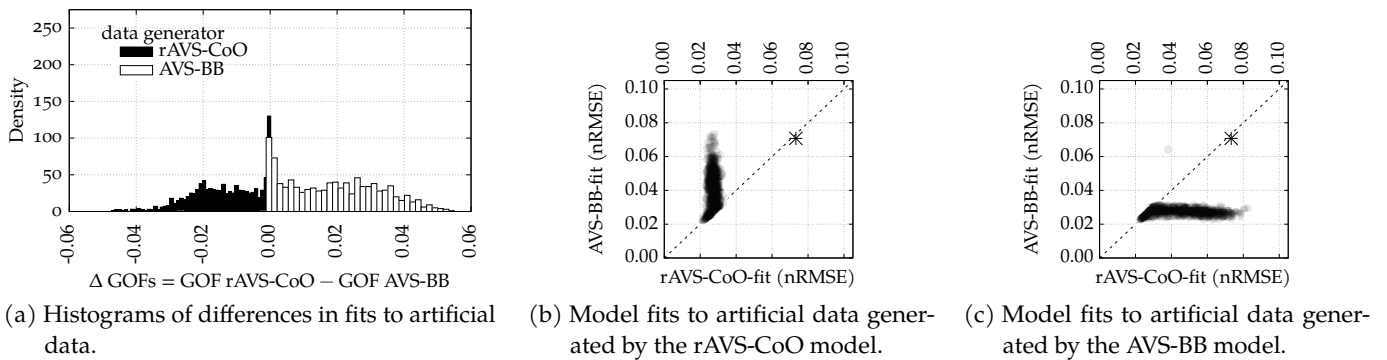


Figure 5.5: Landscaping results contrasting the rAVS-CoO model with the AVS-BB model on the stimuli from Regier and Carlson (2001). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 5.1d). Image copyright: See Appendix E.

flexibility, this finding confirms the same relative ranking of model flexibility for these stimuli as computed by the MFA (AVS-BB slightly more flexible than rAVS-CoO, cf. Figure 5.3 on page 129).

The contrast center-of-object vs. center-of-mass orientation affects model output stronger than the contrasting directionalities of the attentional shift.

RAVS_{W-COMB} VS. RAVS-CoO The two landscaping analyses that contrast the rAVS_{w-comb} model and the rAVS-CoO model are plotted in Figures 5.6 (whole stimuli set) and 5.7 (asymmetrical ROs). The first thing to notice for these results is the overall larger magnitude of model fits compared to the landscaping analyses contrasting the rAVS-CoO and the AVS-BB models (compare axes of Figures 5.6 & 5.7 with Figures 5.4 & 5.5). Apparently, the rAVS_{w-comb} and the rAVS-CoO model mimic each other less than the rAVS-CoO and the AVS-BB model. This is surprising given that conceptually (i.e., in terms of directionality of

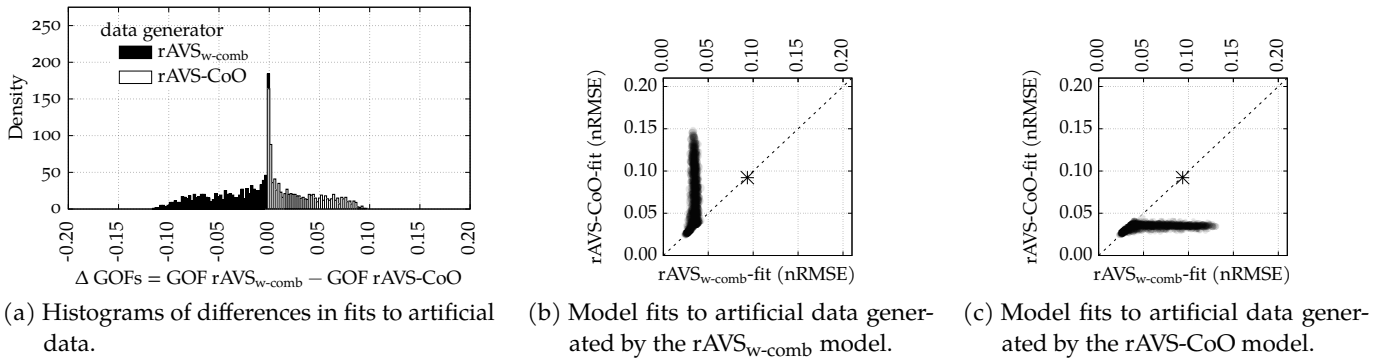


Figure 5.6: Landscaping results contrasting the $rAVS_{w-comb}$ model with the $rAVS-CoO$ model on the whole stimuli set (collapsing across *über*, *above*, and *unter*, *below*). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 5.1a). Image copyright: See Appendix E.

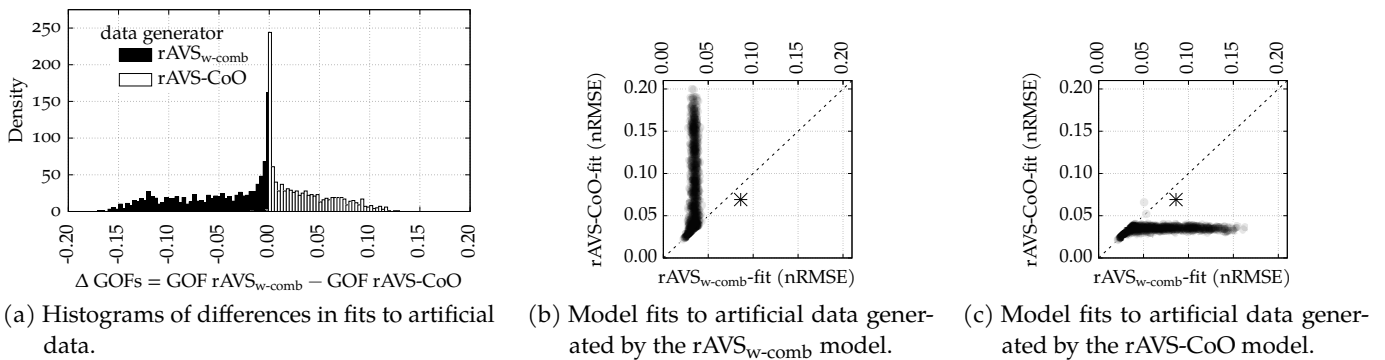


Figure 5.7: Landscaping results contrasting the $rAVS_{w-comb}$ model with the $rAVS-CoO$ model on the asymmetrical ROs only (collapsing across *über*, *above*, and *unter*, *below*). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 5.1c). Image copyright: See Appendix E.

attention), the $rAVS_{w-comb}$ and the $rAVS-CoO$ model are more similar to each other than the $rAVS-CoO$ and the $AVS-BB$ model. This suggests that geometric properties such as the center-of-object orientation are reflected more strongly in acceptability ratings than directionalities of attention.

The reason for comparing the $rAVS_{w-comb}$ model with the $rAVS-CoO$ was the equal model fitting performance of both models for the data from the whole stimuli set from Chapter 4 (cf. asterisks in Panels 5.6b and 5.6c) – despite the superior fit of the $rAVS-CoO$ model to the data from the asymmetrical ROs (cf. asterisks in Panels 5.7b and 5.7c). The landscape results provide evidence that the nature of this equal performance is not due to model mimicry: For both stimuli sets, the two

The $rAVS_{w-comb}$ and the $rAVS-CoO$ model mimic each other only to a very small extent.

models mimic each other only for a small subset of data sets (compare histogram peaks close to zero with long histogram and landscape tails). However, the $rAVS_{w-comb}$ model accounts slightly better for not-self-generated data than the $rAVS-CoO$ model. This is evident from smaller histogram and landscape tails for data generated by the $rAVS-CoO$ model (Panels 5.6c & 5.7c or white histograms in Panels 5.6a & 5.7a) compared to the performance of the $rAVS-CoO$ model for $rAVS_{w-comb}$ -generated data (panels Panels 5.6b & 5.7b and black histograms in Panels 5.6a & 5.7a). This confirms the higher flexibility of the $rAVS_{w-comb}$ model compared to the $rAVS-CoO$ model as computed by the MFA (cf. Figure 5.3 on page 129). The relatively bad performance of the $rAVS-CoO$ model on the whole data set from Chapter 4 seems to be best explained by its missing flexibility to accommodate the effect of relative distance contained in the data from the rectangular ROs.

The $rAVS_{w-comb}$ and the AVS model do not mimic each other.

$rAVS_{w-comb}$ VS. AVS The last landscaping analysis contrasted the AVS and the $rAVS_{w-comb}$ model on the stimuli from Regier and Carlson (2001). The results are plotted in Figure 5.8. Compared to the previous landscaping analyses, these results provide evidence for an even weaker amount of model mimicry (compare height of histogram peaks in Figure 5.8a with the peaks for the other analyses). The landscape plots in Panels 5.8b and 5.8c confirm this finding: The fits are orthogonal to the axis of the data-generating model (i.e., the data-generating model provides relatively constant fits to its own data while the other model varies in its GOF). The landscaping results again confirm the slightly higher flexibility of the AVS vs. the $rAVS_{w-comb}$ model for these stimuli as computed by the MFA results (cf. Figure 5.3 on page 129): Even though the models do not mimic each other, the AVS model fits the $rAVS_{w-comb}$ -generated data slightly better than vice versa (compare length of histogram or landscape tails in Figure 5.8).

Discussion of Landscaping Analyses

Taken together, the landscaping analyses confirmed the relative flexibilities as computed by the MFA. That is, for all model pairs compared with landscaping, the model with the higher MFA-flexibility could fit the not-self-generated data slightly better than the other model (with lower MFA-flexibility). This provides support for the validity of the relative flexibilities computed by the MFA.

In terms of model distinguishability, the landscaping results provide evidence that each model generates data that are different from the data generated by the other model in the respective model pair. This difference is less exaggerated for the $rAVS-CoO$ vs. the AVS-BB model (in particular on the asymmetrical ROs), and relatively strong for the $rAVS_{w-comb}$ vs. the $rAVS-CoO$ model and for the $rAVS_{w-comb}$ vs. the AVS model. Thus, in principle, the models could be distinguished with data collected using these stimuli (in terms of fitting performance).

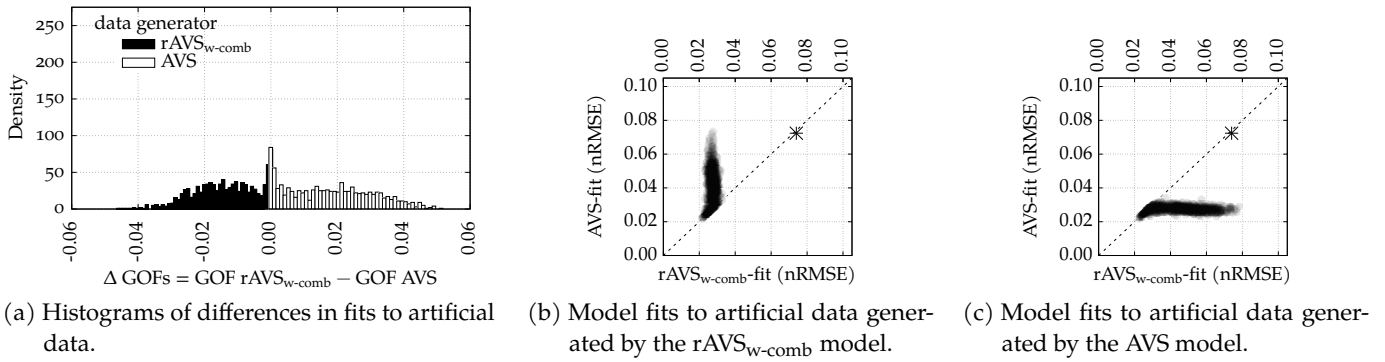


Figure 5.8: Landscaping results contrasting the $r\text{AVS}_{w\text{-comb}}$ model with the AVS model on the stimuli from Regier and Carlson (2001). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 5.1d). Image copyright: See Appendix E.

Why then could I not distinguish the models using the collected empirical data? First, it might be that the landscaping analysis needs more fine-tuning (which in turn affects the interpretation of the results): Potentially, I added too little noise to the artificial data such that the data-generating model was able to fit its own data substantially closer than the possibly more noisy empirical data. This reasoning is suggested by the fact that the fits to the empirical data (asterisks in the landscape plots) are somewhat apart from the fits to the artificial data sets.

Second, it might be that the empirical data are reliably different from the model-generated data (but only to a small extent, see overall magnitudes of model fits). This renders overlapping fits to artificial and empirical data unlikely. In addition, it suggests that all models generate data that are systematically distinct from the empirical data (albeit they are still close, see GOF/SHO). A potential reason for this is the effect of relative distance that neither model appropriately accommodates. Note that this does not mean that the models cannot closely account for the data – they can, see the GOF and SHO results in Figure 5.1. Rather, this means that the not properly captured effect of relative distance might be one reason of not being able to distinguish the models. In other words, the relative distance effect might be a tiger and not a mouse as famously stated by George E. P. Box (1976, p. 792): “Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.”

Finally, comparing the $r\text{AVS}_{w\text{-comb}}$ and $r\text{AVS-CoO}$ models, the comparably low model mimicry surprises. In particular, compared to the other model comparison pairs, some fits of the $r\text{AVS}_{w\text{-comb}}$ and $r\text{AVS-CoO}$ models to artificial data are twice as high (i.e., worse). This is surprising because the $r\text{AVS}_{w\text{-comb}}$ and $r\text{AVS-CoO}$ model are conceptually close in terms of the implemented directionality of the attentional

shift (from the LO to the RO). On the other hand, the other model comparison pairs contrasted the two directionalities of the shift but revealed closer fits to the generated artificial data. This suggests that the contrast center-of-object orientation vs. center-of-mass orientation – the dimension that differs between the $rAVS_{w-comb}$ and $rAVS-CoO$ model – affects the generated model output stronger than the different directionalities of the attentional shift.

5.6 DISCUSSION OF ALL MODEL SIMULATIONS

Considering all model simulations, the two newly proposed models $rAVS-CoO$ and $AVS-BB$ (accounting for the center-of-object orientation instead of the center-of-mass orientation) perform substantially better than their predecessors $rAVS_{w-comb}$ and AVS . In contrast to the center-of-mass models, the two center-of-object models better fit the empirical data (GOF, SHO) while they are less flexible (MFA, landscaping) and generate rating patterns closer to the empirical patterns (PSP). This supports the idea that people rely on the center-of-object orientation instead of on the center-of-mass orientation.

While this is an interesting finding, it does not answer my main research question whether attention shifts from the RO to the LO (AVS , $AVS-BB$) or from the LO to the RO ($rAVS_{w-comb}$, $rAVS-CoO$). This is because I still cannot reliably distinguish these two model classes in terms of performance on empirical data. Both implemented directionalities of attention account equally well for the empirical data. On the other hand, the clearly superior performance of the center-of-object models ($AVS-BB$, $rAVS-CoO$) vs. the center-of-mass models (AVS , $rAVS_{w-comb}$) suggests that geometric properties such as center-of-object vs. center-of-mass orientation are more important for model performance than the implemented directionality of the attentional shift.

A likely reason for a strong effect of geometry is the sole evaluation of model goodness using acceptability ratings. This kind of empirical data obviously does not measure shifts of attention. Ratings are only indirectly influenced by attentional shifts via the perceptual processing of geometric properties. Future research should extend the models to more specifically implement visual attention, e.g., by directly modeling eye movements (as a measure of shifts of overt attention) or integrating a temporal component (because attentional shifts are inherently temporal). I present related ideas in more detail in Chapter 6. As a first step toward enhancing the models to generate data that can be better matched to human responses, the next and final section of the present chapter introduces an extension that enables the models to simulate rating distributions instead of mean ratings.

The center-of-object models outperform the center-of-mass models on all measures.

Based on simulating contrasting implementations, the two directionalities of attention cannot be teased apart. Both implementations account equally well for the empirical data.

5.7 OUTLOOK: RATING DISTRIBUTIONS AND BAYESIAN INFERENCE

I developed the following model extension because the models are only evaluated using mean ratings. However, by summarizing rating distributions to a single mean rating, the richness of rating data is ignored. In addition, one should not treat rating data as metric (e.g, by computing a mean rating, Liddell & Kruschke, 2018). Accordingly, I extended the models such that they are considering rating data properly as ordinal (discrete and ordered) data by simulating a rating distribution. Section 5.7.1 presents this model extension in detail.

Apart from a more correct way of handling the data, the model extension further allows to generate individual ratings by sampling from the simulated rating distribution. Using the cross-match test proposed by Rosenbaum (2005) as likelihood function, this allows to apply Bayesian inference for the model parameters. In Section 5.7.2, I introduce this method in detail and present an example application.

Due to time constraints, I could evaluate the model extension and the application of the cross-match test only in a limited way (only for the $rAVS_{w-comb}$ and $rAVS-CoO$ model and the data from the asymmetrical ROs). That said, I believe that the proposed methods create a wide range of opportunities to further explicate and empirically test theoretical claims of AVS-like models. I hope that this initial work inspires other researchers to pursue this path.

5.7.1 *Rating Distributions*

The model extension to simulate rating distributions was inspired by the method with which I analyzed the empirical rating data (see Section 4.2.1), namely ordinal regression models (Kruschke, 2015, Chapter 23, Liddell & Kruschke, 2018). These regression models assume a latent metric distribution underlying the ordinal distribution. The probability of a single rating is then defined as the cumulative probability between two thresholds defined on the metric distribution.

As an example, consider Figures 5.9c and 5.9d. In these figures, the two parts of the ordinal regression are plotted as dashed lines: The latent metric distribution is a dashed curved line, the thresholds are dash-dotted vertical lines. In addition, the cumulative probabilities between two thresholds are plotted as crosses (computed with the extended $rAVS-CoO$ model, Figure 5.9c) or asterisks (computed with the extended $rAVS_{w-comb}$ model, Figure 5.9d). Based on ordinal regression models, the proposed model extension works as follows:

* The work reported in Section 5.7 (published as Kluth & Schultheis, 2018) profited from ideas and feedback from Holger Schultheis. In particular, the idea to use the cross-match test originates from him.

1. treat the outcome of an AVS-like model (mean rating) as the mean μ of a Gaussian distribution; add the standard deviation σ of the Gaussian distribution as an additional model parameter
2. define $K - 1 - 2$ thresholds on the Gaussian distribution and add them as additional model parameters; K is the number of ratings on the ratings scale; the first and last threshold are fixed (half-way between the values of the first and second rating or last and second-last rating)
3. compute the probabilities of each rating as the cumulative probability between two thresholds (or between the first/last fixed threshold and negative/positive infinity for the first/last rating)

In Figures 5.9c and 5.9d, two empirical rating distributions from the empirical study presented in Chapter 4 are plotted as bars, exemplarily for the left and right LO above the asymmetrical RO shown in Figure 5.9a. Clearly, the study participants rated the right LO (more central with respect to the center-of-object) higher than the left LO (less central). More precisely, they picked the rating 9 more often for the right LO than for the left LO (cf. to greater empirical data set and statistical model fits plotted in Figure 4.13 on page 97).

Compared to evaluating the cognitive models on mean ratings, the model extension makes a more fine-grained model evaluation possible. In Figures 5.9c and 5.9d, the probabilities for each rating as computed with the extended $rAVS_{w-comb}$ and $rAVS-CoO$ models are plotted (from now on and in Figure 5.9, extended models are denoted with a trailing +). While both models account relatively well for the empirical ratings for the right LO (Figure 5.9d), the $rAVS_{w-comb+}$ model clearly over-estimates the probability for rating 9 for the left LO (Figure 5.9c). Due to equal center-of-mass orientation for both LOs, the $rAVS_{w-comb+}$ produces exactly the same outcomes for both LOs. This comes as no surprise (considering the mechanism of the $rAVS_{w-comb}$ model) but it does not capture the empirical data.

Compare this model evaluation using rating distributions with a model evaluation using mean ratings (for the left LO 7.38, for the right LO 8.18). With the same parameter settings, this yields a model fit for the left LO of 0.1326 ($rAVS_{w-comb}$ fit, nRMSE) or 0.0093 ($rAVS-CoO$ fit, nRMSE) and for the right LO 0.0333 ($rAVS_{w-comb}$ fit, nRMSE) or 0.1029 ($rAVS-CoO$ fit, nRMSE). None of these numbers provides information about the model properties as intuitive and informative as the fit of the extended models using full rating distributions.

To more systematically assess the simulated rating distributions with respect to empirical rating distributions (relative frequencies of ratings),

Simulating full rating distributions allows for a more fine-grained model evaluation (compared to using mean ratings).

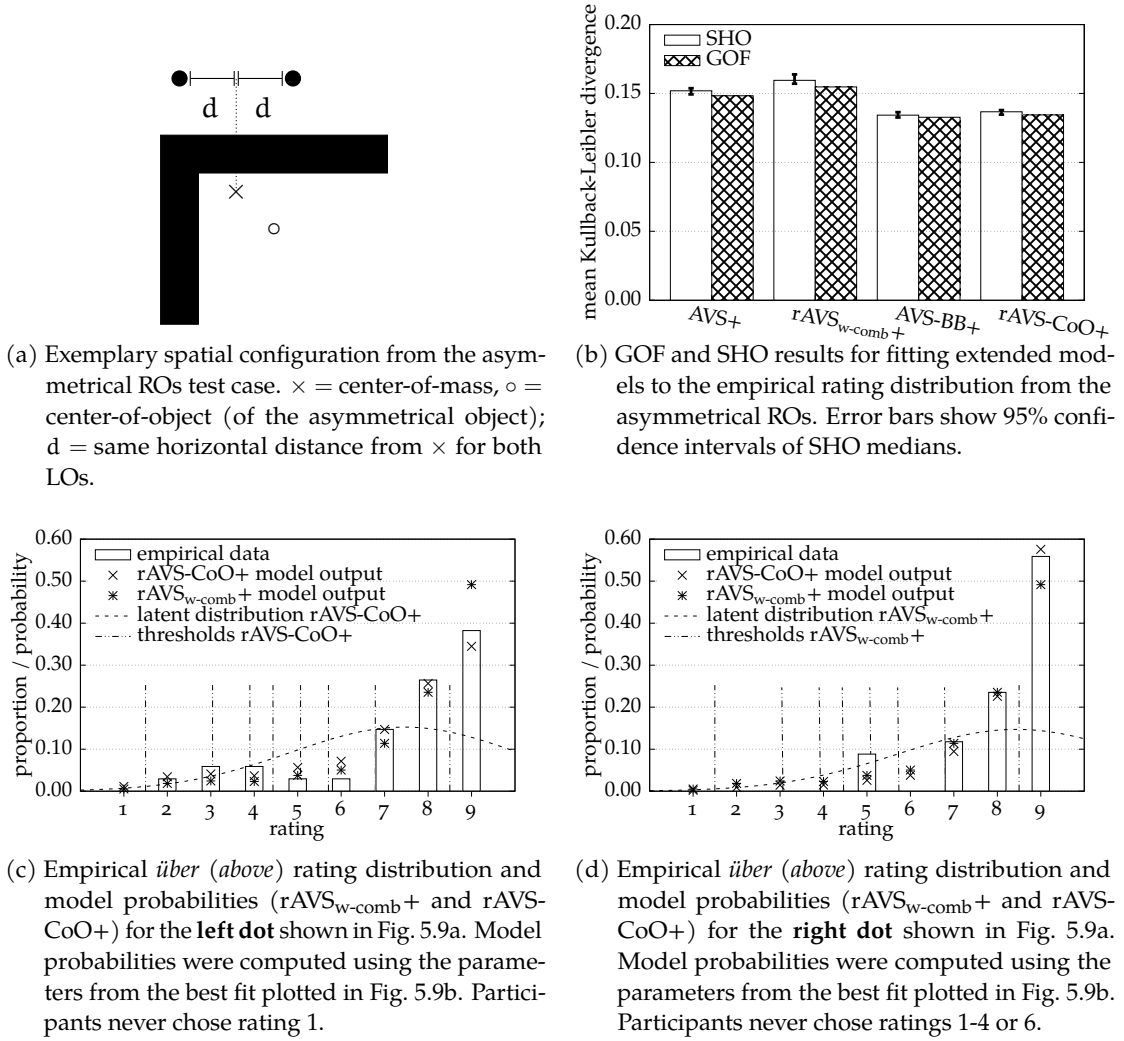


Figure 5.9: Example experimental display to illustrate the model extension that enables the simulation of rating distributions, fits of extended models, and empirical rating distributions. Image copyright: See Appendix E.

I propose to use the Kullback-Leibler (KL) divergence¹⁰ from the simulated rating distribution (P_{sim}) to the empirical rating distribution (P_{emp}). The KL divergence is defined as:

$$D_{KL}(P_{emp}||P_{sim}) = - \sum_{i=1}^K P_{emp}(i) \log \frac{P_{sim}(i)}{P_{emp}(i)} \quad (5.7)$$

The KL divergence is computed for each RO-LO pair. To find the best fitting parameters, I minimized the mean KL divergence (averaged over all stimuli).

¹⁰ Solomun Kullback preferred the term “discrimination information” (Kullback, 1987). Despite this, the measure is widely known as Kullback-Leibler divergence.

Simulating full rating distributions yields the same relative model performances as considering mean ratings only.

To test the proposed extension, I used the data set from the asymmetrical ROs test case. That is, I implemented the model extension for the AVS, $rAVS_{w-comb}$, AVS-BB, and $rAVS-CoO$ models and fitted (GOF & SHO) the extended models by minimizing the mean Kullback-Leibler divergence. The results can be seen in Figure 5.9b. Comparing this plot with Figure 5.1c (page 119), confirms the validity of the model extension as the general pattern of relative model performance is reproduced (center-of-object models fit the data better than center-of-mass models).

The model extension also allows the models to generate individual data by sampling from the simulated rating distribution. Note that the simulated rating distribution is completely determined by the model parameters (i.e., it does not change without changing the model parameters) while the sampled individual ratings are subject to sampling noise. Apart from modeling individual behavior (a task that goes well beyond my work, see Navarro, Griffiths, Steyvers, & Lee, 2006, for useful ideas), the individual ratings can be used as input for the next proposed method that enables Bayesian inference for AVS-like models.

5.7.2 Bayesian Inference Using the Cross-Match Test

One strength of Bayesian data analysis is that it allows to reason about the values of model parameters by using full probability distributions (the posterior distributions; cf. my analyses of the empirical data in Section 4.2). For regression analyses, this means to quantify the effect size (mode of posterior distribution) as well as the uncertainty of the model (spread of the posterior distribution). The same principle, Bayesian inference, is also applicable to model parameters from cognitive models. Different from statistical models, the parameters of cognitive models are often linked to assumed cognitive representations or mechanisms (e.g., λ in the AVS model as attentional width). The likely values of these parameters are informative to further understand how a model works and to inform cognitive theorizing.

Note that while Bayesian inference is natural for so-called Bayesian models of cognition (e.g., Chater, Oaksford, Hahn, & Heit, 2010), these models also assume that (parts of) cognition can be best described by using probability theory (because, so the argument goes, humans have to deal with uncertainty almost everywhere). It is debated whether this assumption holds true (e.g., Jones & Love, 2011). Crucially, however, it is technically not necessary to make this assumption for the application of Bayesian inference for cognitive models.

What is necessary for Bayesian inference, though, is a likelihood function. This function specifies how likely empirical data are given a specific parameter set. All AVS-like models lack a likelihood function. In the following, I propose to use the cross-match test by Rosenbaum (2005) as an approximation of the likelihood function.

Cross-Match Test

The cross-match test (Rosenbaum, 2005) is a statistical test that computes the probability whether multivariate responses from two groups come from the same distribution. In my case, a multivariate response (rows in Table 5.1) consists of ratings to all stimuli. These responses can stem from two groups: human participants or model simulations. The cross-match test computes rank-based Mahalanobis distances between all responses and matches the closest responses. If two responses from different groups are matched, this is called a “cross-match”. The more cross-matches exist, the more likely it is that the data in both groups come from the same distribution (see Rosenbaum, 2005, for more details).

I cannot change the empirical data but I can change the model-generated data (by choosing different model parameters). This makes the cross-match test a likelihood function measuring how likely the model-generated data (dependent on model parameters) come from the same distribution as the human data.

Method: Computing the Likelihood Function

The cross-match test requires that a model generates individual data. The model extension that simulates rating distributions enables such generation of individual data. To produce one “artificial participant”, the model simulates the rating distributions for the corresponding stimuli and afterwards one can sample individual ratings from these rating distributions. To compute the likelihood functions of the rAVS_{w-comb}+ and the rAVS-CoO+ models, I generated as many artificial subjects as I tested human participants (34). Then, I computed the cross-match test comparing the 34 model-generated data sets with the 34 human data sets.

The sampling from the rating distributions involves sampling noise, i.e., the same model parameters will lead to different data sets. However, the likelihood function should be approximately stable for the same parameters. This is why I had to follow a more sophisticated process: For every artificial rating, I sampled s times from the corresponding rating distribution and used the mean rating (rounded to be an integer) as generated rating. To further stabilize the outcome of the cross-match test, I computed an average of several iterations of data generation and cross-match tests. More precisely, I computed the mean number of cross-matches for c iterations and stored the resulting probability for this number of cross-matches. I did this for b blocks with the final likelihood value being the mean probability of the b single probabilities. Taken together, the computation of a single likelihood value requires the generation of $34 \times s \times c \times b$ individual data sets (with 34 human data sets). I obtained a relatively stable likelihood value with $s = 10$, $c = 4$, $b = 20$ (standard error of averaged cross-match results < 0.05).

Table 5.1: Example input for the cross-match test (Rosenbaum, 2005). Each row describes the response of one subject (empirical or model-generated), each column describes the response to a stimulus (e.g., the left or right LO from Fig. 5.9a). Table copyright: See Appendix E.

data type	left LO	right LO	...
empirical	7	8	...
empirical	9	9	...
...
model	8	9	...
model	5	8	...
...

Method: Estimating the Posterior Distribution

The posterior distribution of the model parameters is determined by empirical data (via the likelihood function) and the prior distribution over the parameter ranges. These prior distributions should consider previous knowledge about the likely values of the model parameters. Since this is the first study with probabilistic AVS-like models, I decided to use “uninformative” prior distributions defined as uniform distributions over the ranges of the parameters (see Equations 3.15–3.17, page 42, and 5.1–5.2, page 118).

To test the cross-match approach, I estimated the posterior distribution of the rAVS-CoO+ model for the data from the asymmetrical ROs. First, I estimated the posterior distribution also for the additional model parameters for the model extension (σ of latent Gaussian distribution and thresholds). Due to the larger parameter space, the additional model parameters complicated the convergence process of the posterior estimation (taking more time and computational resources), while – compared to keeping the extension parameters constant – the qualitative results for the four original model parameters (α , highgain, intercept, slope) were not affected. This is why I kept the extension parameters constant on the values of the best rAVS-CoO+ fit to the data from the asymmetrical ROs.

Given a model with a likelihood function, prior distributions, and an empirical data set, one can apply standard ‘Markov Chain Monte Carlo’ (MCMC) methods to estimate the posterior distribution. To do so, I extended the Metropolis-Hastings algorithm (already implemented for parameter fitting, see Algorithm 1 in Section 3.2.1). Instead of searching for one best parameter set that minimizes the nRMSE or the mean KL divergence, the MCMC algorithm estimates the posterior distribution by visiting parameter sets θ proportionally to the corresponding posterior value (with $\text{posterior}(\theta) \sim \text{prior}(\theta) \cdot \text{likelihood}(\theta)$). I estimated the posterior distributions with four MCMC chains (125,000

samples each) and checked convergence with the potential scale reduction factor \hat{R} (Gelman & Rubin, 1992). To improve the convergence of the MCMC algorithm, I implemented the adaption algorithm by Garthwaite, Fan, and Sisson (2016). For the cross-match test, I used the R package `crossmatch` (Heller, Small, & Rosenbaum, 2012) and re-implemented parts of it using the C++ library `Armadillo` (Sanderson & Curtin, 2016). The R package `ggmcmc` (Fernández-i-Marín, 2016) helped in visualizing and analyzing the MCMC samples. All source code is available from Kluth (2018).

Results: Bayesian Inference

The marginal posterior distributions are visualized in Figure 5.10. Each chain is plotted with a different color. The overlap of the different chains confirms the convergence of all MCMC chains. At a first glance, the marginal posterior distributions are surprising because they do not have clear modes. Rather, almost all parameter values seem to be equally likely with respect to the empirical data. In particular the `highgain` parameter and the α parameter have relatively flat distributions, while the profiles for the intercept and slope parameters are more diverse.

A high posterior density for a certain parameter range means that these parameter values are more likely than other parameter values. Since posterior distributions are based on empirical data one can conclude the following: Using such more likely parameter values, the model performance in terms of fitting data should be better compared to using less likely parameter values. Applying this conclusion to double-check the validity of the unexpected posterior distributions, I picked two parameter sets based on the maxima/minima of the distributions. According to the posterior distributions, the first parameter set is supposed to result in bad model performance ($\alpha = 0.2$, `highgain` = 5.0, `intercept` = 1.25, `slope` = -0.05). In contrast, the second parameter set ($\alpha = 3.0$, `highgain` = 5.0, `intercept` = 0.9, `slope` = -0.625) should perform relatively well.

These expectations are confirmed by two independent goodness-of-fit measures. Using the extended model, the first parameter set fits the empirical data worse than the second (mean Kullback-Leibler divergence: 0.484 against 0.266, respectively). Also, for the non-extended `rAVS-CoO` model, the first parameter set is worse than the second (nRMSE for worse parameters 0.301 against 0.145 for better parameters). These GOFs provide support for the unexpected shape of the marginal posterior distributions.

What do the marginal posterior distributions in Figure 5.10 now tell us? First, the parameter `highgain` seems to be irrelevant for model performance as its posterior distribution has a flat profile throughout the parameter range – all values of `highgain` are equally likely. Next, low values of the α parameter correlate with bad model performance (the posterior density of α decreases on the left hand side). Given that

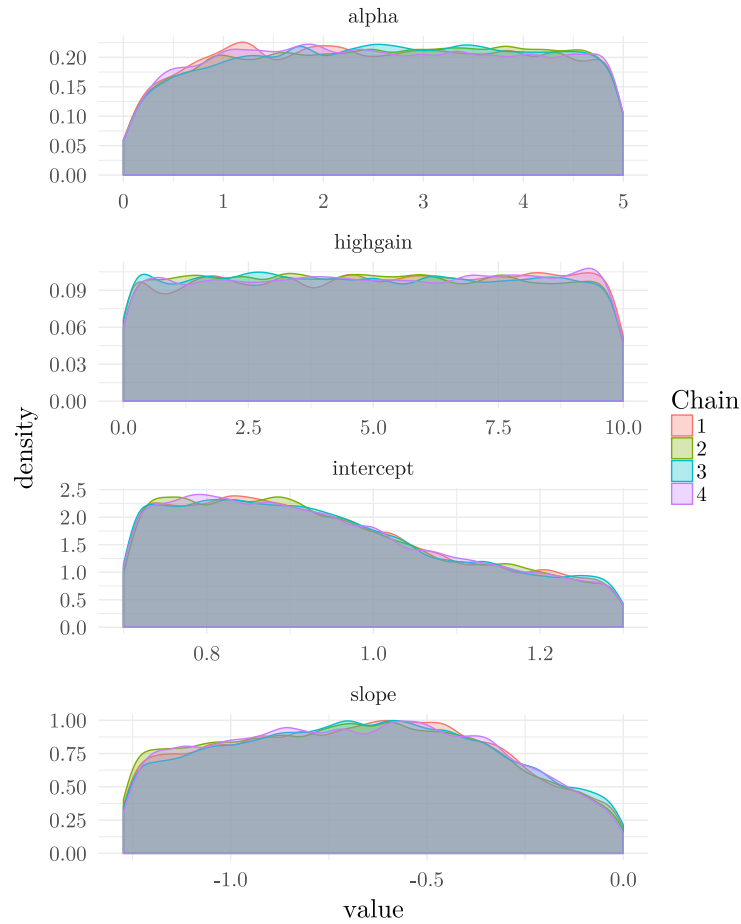


Figure 5.10: Marginal posterior distributions for the rAVS-CoO+ model given rating data from asymmetrical ROs and “uninformative” prior distributions (uniform distributions). Image copyright: See Appendix E.

α weights the importance of the proximal vs. center-of-object orientation in the rAVS-CoO+ model, this means that the center-of-object orientation seems to be more important than the proximal orientation for accommodating the asymmetrical RO data. This is because lower values of α that favor the proximal orientation over the center-of-object orientation result in a lower value of the posterior distribution – i.e., lower values of α (advantage of proximal orientation) are less likely than higher values of α (advantage of center-of-object orientation).

The two parameters α and highgain are part of the model component that processes the spatial configuration and the geometry of the RO. The outcome of this model component is an angular deviation from a reference direction. A second model component maps this angular deviation to an acceptability rating. This second model component consists of a linear function and relies on the two parameters intercept and slope. The marginal posterior distributions for intercept and

slope show more diverse profiles with relatively clear maxima/minima compared to the α and highgain distributions. When comparing the effects on model performance by changing the values of intercept and slope to the same degree as α and highgain (relative to the allowed parameter ranges), changing intercept or slope will most likely affect model performance stronger than changing alpha or highgain. These observations are somewhat qualified considering the decrease of the posterior density of α for small values.

Nonetheless, an interesting conclusion from the posterior distributions is that the second component (mapping angular deviation to ratings using the parameters intercept and slope) seems to have a greater influence on model performance than the first component (dealing with geometric processing using the parameters α and highgain). This conclusion is contrary to the fact that most researchers using experimental paradigms similar to the study presented in Chapter 4 are primarily interested in the question how the processing of geometric properties of depicted spatial relations affects spatial language evaluation. In contrast, my results suggest that the question of how the outcome of perceptual processes is mapped to linguistic judgments deserves more attention than the perceptual processing itself. Note that all conclusions are only valid for the tested model and data set. Future research should investigate them more closely.

Moreover, Vishal Singh (Indian Institute of Technology) conducted unpublished follow-up work¹¹ to validate the approach of using the cross-match test as a likelihood approximation. This follow-up work casts serious doubt on the validity of the cross-match approach, as the cross-match likelihood failed to estimate known posterior distributions. In terms of the above conclusion (mapping component more important than geometry processing component), this calls for further investigations using different tools. Certainly a fruitful choice for this is the ‘Approximate Bayesian Computation’ toolkit (ABC, see Palestro, Sederberg, Osth, van Zandt, & Turner, 2018; Turner & Van Zandt, 2012, for general introductions; Turner & Sederberg, 2014, for a promising algorithm). The goal of ABC is to provide tools that enable cognitive modelers to analyze non-probabilistic cognitive models with Bayesian methods. Because many cognitive models lack a likelihood function, ABC algorithms provide so-called “likelihood-free Bayesian analysis”. Unfortunately, time did not allow to apply ABC methods for AVS-like models in this Ph.D. project.

Mapping angular deviation to a rating seems to affect model performance stronger than computing the angular deviation from geometry.

11 in the scope of an internship at the University of Bremen under the supervision of Holger Schultheis

Part III

GENERAL DISCUSSION

TOWARDS A COMPREHENSIVE MODEL OF SPATIAL LANGUAGE PROCESSING

The computational and empirical studies conducted within this Ph.D. project were motivated by the following question: In order to process the description of a spatial relation, are humans shifting their attention from the RO to the LO or from the LO to the RO? The main outcome of this project is that one cannot reliably answer this research question – given the existing empirical data and the implementations of directed attentional shifts (AVS, AVS-BB, rAVS_{w-comb}, and rAVS-CoO models).

This is why, in Section 6.2, I discuss future directions (based on the findings from the present research) on how to enhance the cognitive models in order to obtain a more comprehensive model of spatial language processing. Such a model might provide more fine-grained answers to the question of the directionality of the attentional shift. In order to relate my findings to future model enhancements, I follow the seminal three-level approach proposed by David Marr (1982). Doing so, connects the currently binary question (shift from RO to LO vs. LO to RO; cf. Newell, 1973) to further relevant research. This allows cognitive scientists to ask more subtle questions in the future and thus to more fully understand spatial language processing. First, this chapter starts in Section 6.1 with a summary of the results described in the previous chapters. In Section 6.4, the thesis closes with a final conclusion.

6.1 SUMMARY OF FINDINGS

In Chapters 1 and 2, this thesis started with a review of literature relevant for understanding the contribution of visual attention to spatial language verification. From this review, one important insight emerged: While shifts of attention seem to be an inherent part of processing spatial relations (Franconeri et al., 2012), the specific direction of such shifts for the processing of spatial prepositions remains unclear. On the one hand, influential theoretical research on spatial language processing claims that people should shift their attention from the RO to the LO (Logan, 1994, 1995; Logan & Sadler, 1996; Logan & Zbrodoff, 1999). On the other hand, more recent empirical studies suggest that attention moves from the LO to the RO (Burigo & Knoeferle, 2015; Roth & Franconeri, 2012). The AVS model (Regier & Carlson, 2001) can be interpreted as an implementation of the “traditional” account of a shift from the RO to the LO. Hence, this Ph.D. project set out to modify the AVS model in order to implement a reversed shift from the LO to the RO. I called this modification the reversed AVS (rAVS) model. If any of

these two implementations (that differ in their directionality of attention) performs better on human data than the other, this would provide evidence in favor of the superiority of the implemented directionality.

In Chapter 3, I developed several model variations that reverse the attentional shift in the AVS model. These rAVS variations were based on two geometrical properties known to affect acceptability ratings: the proximal orientation and the center-of-mass orientation. I assessed all rAVS variations on the empirical data collected by Regier and Carlson (2001, *above* acceptability rating data). One rAVS variation – the rAVS_{w-comb} model – performed equally well on these data as the AVS model. Thus, both directionalities of attention seemed to be equally likely to describe these human data.

Chapter 4 presented novel stimuli with the goal of contrasting the two shift-implementations regarding their distinct predictions for these stimuli. More specifically, I created two test cases: a relative distance test case (consisting of rectangular ROs with different heights) and an asymmetrical ROs test case (investigating the influence of asymmetrical mass distributions on ratings). I conducted an empirical rating study that tested whether humans followed the model predictions. The rating study revealed that the relative distance of the LO to the RO (i.e., absolute distance divided by width and height of the RO, see Equation 3.5 on page 37) affected acceptability ratings. Regression analyses showed that relative distance modulated the effects of the proximal orientation and the center-of-mass orientation: Higher relative distance strengthened a *reversed* effect of center-of-mass orientation and lower relative distance weakened the effect of proximal orientation. While this finding confirms the general prediction of the rAVS_{w-comb} model (relative distance should modulate the effect of the two orientations), the rAVS_{w-comb} model cannot account for the specific qualitative interactions of the predictors relative distance, proximal orientation, and center-of-mass orientation. The AVS model also fails to fully accommodate the empirical results. Future research should investigate the relative distance effect more closely.

Relative distance modulates the effects of the proximal orientation and the center-of-mass orientation.

For the asymmetrical ROs, the empirical results suggest that people rather rely on the center-of-object orientation than on the center-of-mass orientation. This finding goes against the importance of the center-of-mass orientation implemented in both the AVS and the rAVS_{w-comb} model. To account for the seemingly greater importance of the center-of-object orientation, I developed modifications to both models (the AVS-BB model and the rAVS-CoO model) and presented these in Chapter 5. While these center-of-object models outperform their predecessors, future research should analyze to what extent a flat vs. non-flat top/bottom of the RO (facing the LO) qualifies the importance of the center-of-object vs. center-of-mass orientation.

People seem to prefer the center-of-object orientation over the center-of-mass orientation.

In terms of eye movements (overt attention), the empirical data confirmed the horizontal component of the attentional focus as defined in

the AVS model. The $rAVS_{w-comb}$ model also uses this point. However, the design of the study does not allow me to interpret the data as confirming the vertical component of the hypothesized focus. Although the empirical gaze patterns do not contradict the theorized vertical location of the focus, one cannot say whether this is due to the used preposition or the vertical location of the LO. For the asymmetrical ROs, the study revealed two different gaze patterns: While for the more open asymmetrical ROs (L-shaped), participants' fixations were influenced by the asymmetrical mass distribution, the gaze patterns for the more closed asymmetrical ROs (C-shaped) could not be distinguished from fixation patterns to the rectangular ROs.

Chapter 5 presented the outcomes of several model evaluation techniques using data and stimuli from the conducted empirical study. These model simulations revealed that, in contrast to the center-of-mass models AVS and $rAVS_{w-comb}$, the two newly proposed center-of-object models AVS-BB and $rAVS-CoO$ better fit the empirical data (GOF and SHO) while they are less flexible (MFA, landscaping) and generate rating patterns closer to the empirical patterns (PSP). This supports the idea that people rely on the center-of-object orientation instead of on the center-of-mass orientation. However, in terms of distinguishing between the two contrasting shifts of attention – from the RO to the LO, AVS and AVS-BB, or from the LO to the RO, $rAVS_{w-comb}$ and $rAVS-CoO$ – the simulation results remain inconclusive. Neither implemented directionality outperforms the other. One potential reason for this is the failure of both models to capture the effect of relative distance. Taken together, the simulation results highlight the relatively strong effects of geometrical properties on spatial language verification (relative distance, center-of-object orientation). By contrast, the effect of reversing the directionality of the attentional shift on acceptability ratings seems to be too indirect to be reflected in distinct model performances.

In order to investigate the models with a greater level of detail, Chapter 5 closed with presenting a model extension. This extension enables AVS-like models to generate full rating distributions instead of mean ratings. This allows to use all available information in the empirical data to assess the models (because data aggregation such as averaging is not needed). The remainder of the present chapter presents ideas for further model refinements that could help to untangle the role of the directionality of the attentional shift. To do so, I reconsider related research from Chapters 1 and 2 as well as discuss the just summarized findings from this Ph.D. project. The seminal three-level framework proposed by Marr (1982) serves as a structuring guideline.

The center-of-object models outperform the center-of-mass models.

Neither directionality of attention accounts better for the empirical data than the other.

6.2 LEVELS OF ANALYSIS: MARR'S THREE-LEVEL PROPOSAL

More than 30 years after its publication, Marr's three-level suggestion remains relevant for cognitive science (e.g., reflected by the re-publication

*computational: what
and why?
algorithmic: how?
(algorithms &
representations)
implementational:
how? (neuronally)*

of his book *Vision*, Marr, 2010, or the recent special issue in *Topics in Cognitive Science*, Peebles & Cooper, 2015). Although originally proposed for computational models of human vision, nowadays Marr's three levels are considered useful for all domains of cognitive modeling. Any cognitive model can and should be analyzed on three interacting levels: the computational level, the algorithmic/representational level, and the implementational level. Each level is subject to different questions regarding the modeled task. The computational level asks what the cognitive system computes – and why. The algorithmic/representational level specifies how the task is computed in terms of algorithms (or processes) and representations. Finally, the implementational level asks how the algorithms and representations might be implemented in the brain. How could one interpret AVS-like models using these three levels?

6.2.1 AVS-like Models and Marr's Levels

COMPUTATIONAL LEVEL The answer to the “what is computed?” question of the computational level is straight-forward: a linguistic acceptability rating of a spatial preposition, given a depicted spatial relation. Why would humans compute such a rating? The main motivation for the development of AVS-like models was to investigate how people ground (spatial) language in the external world. That is, despite the infinite number of possible spatial configurations, human language parses space in few (relatively distinct) spatial categories and uses spatial prepositions to describe these categories. A reason for this categorization might be that a central purpose of language is efficient communication. Typically, there is no need to know the exact location of an object. Accordingly, language users can rely on broad but flexible categories (via spatial relations) instead of having to negotiate more detailed spatial aspects of their utterances. However, the membership of a spatial relation to a linguistic spatial category must be somehow computed. A linguistic acceptability rating measures this membership.¹

ALGORITHMIC/REPRESENTATIONAL LEVEL I claim that AVS-like models are primarily specified on the algorithmic/representational level. That is, they try to answer how (in terms of algorithms/processes and representations) humans compute linguistic acceptability ratings of spatial prepositions. More specifically, AVS-like models assume (i) polygons as (perceptual) representations of the two objects that are part of the relation, (ii) labels as (linguistic) representations that distinguish the RO from the LO, (iii) vectors as (perceptual) representations of the spatial relation, and (iv) a canonical direction as (linguistic?)

¹ It remains an interesting question whether humans actually compute such ratings (or the like) during spatial language processing or whether the ratings are merely an artifact of psycholinguistic experiments.

representation of the prototypical meaning of the to-be-judged spatial preposition.

In terms of algorithms or processes, AVS-like models implement (i) an attentional selection of one of the two objects (with their attentional distribution, see also Section 2.3), (ii) an attentional shift from one object to the other that yields a direction, (iii) an angular comparison of directions to obtain an angular deviation, and (iv) a mapping of angular deviation to linguistic acceptability rating. Different variations of the models further implement specific mechanisms that process the spatial properties of the involved objects: In the AVS and AVS-BB models, the weighted vector sum (a type of spatial pooling, see page 20) translates the spatial relation into the attentional shift; in the $rAVS_{w-comb}$ and $rAVS-CoO$ models, relative distance, proximal orientation, and center-of-mass/object orientation interact with each other to yield the attentional shift.

IMPLEMENTATIONAL LEVEL On the implementational level (how are the proposed mechanisms and representations from the algorithmic/representational level implemented in the brain), AVS-like models are underspecified. Admittedly, Regier and Carlson (2001) motivate the weighted vector sum with neuroscientific research on neuronal population codes for movements (Georgopoulos et al., 1986; Lee et al., 1988; Wilson & Kim, 1994). Additionally, the attentional distribution certainly could be linked to respective neuroscientific research on spatial attention. However, it seems more difficult to link other core model parts to potential neurological substrates (e.g., comparison of angles or mapping of angular deviation to linguistic rating).

Based on this interpretation of AVS-like models in terms of Marr's three levels and on the literature reviewed in Chapters 1 and 2, the remainder of this chapter conceptualizes both the findings of this Ph.D. project as well as potentially fruitful model extensions following Marr's framework.

6.2.2 *Extending the Computational Level: The Role of Language in (Spatial) Category Perception*

To describe cognition on the computational level, researchers have used Bayesian models of cognition that often describe a rational or optimal solution to problems faced by cognitive systems (see e.g., Chater et al., 2010; Griffiths, Kemp, & Tenenbaum, 2008, for reviews). Although this general endeavor has been criticized (Jones & Love, 2011, in particular for assuming that cognition aims for (mathematically) optimal solutions using rational principles), I think it is a promising way to spell out the computational level of AVS-like models in more detail.

In the domain of space, language applies spatial categories. One important goal of language use is successful communication with others.

More specifically, spatial language should help listeners to find objects. Given a spatial sentence and a scene, the spatial description helps in narrowing down the search space, i.e., where to look for the located object. The more the LO position overlaps with the intended spatial category, the higher the probability of efficiently finding the LO.

Phrased in Bayesian terms, spatial language processing should maximize the (posterior) probability of finding the LO based on a prior known spatial category given the actual location. Relating a spatial description to a depicted spatial relation then boils down to compute (a graded) membership of the depicted relation to the linguistic category (used in the spatial description). The acceptability rating computed by AVS-like models can be interpreted as such a graded membership value.

Interpreted in this way, the task of relating (categorical) spatial prepositions to (fine-grained) spatial relations reminds of the distinction between categorical vs. coordinate spatial relation processing from cognitive neuroscience research (see Section 2.2.2). This research revealed that visual attention plays an important role for processing a spatial relation as either categorical or coordinate. More specifically, categorical processing of spatial relations is enhanced, if people use a small attention window (only selecting one object of the relation at a time). By contrast, with a large attention window, coordinate processing of spatial relations improves. This supports the general claim that shifts of visual attention are crucial for categorical spatial relations (see also the discussion of this *process* in Section 6.2.3).

While missing the explicit link to the neuroscience literature, the Category Adjustment (CA) model proposed by Huttenlocher, Hedges, and Vevea (2000, see also Crawford, Huttenlocher, & Hedges, 2006) distinguishes between prior categorical information and fine-grained encoded stimuli, too. The CA model holds that for stimulus judgments (not restricted to spatial judgments), humans apply Bayesian estimation to maximize the average accuracy. More precisely, the CA model assumes that prior categorical information affects the re-production of fine-grained but inexact encoded stimuli. One prediction from the CA model is that stimuli reproduction is biased towards the center of categories. Huttenlocher, Hedges, Corrigan, and Crawford (2004) tested this prediction with spatial relations. The results from Huttenlocher et al. (2004) are summarized in Section 6.2.3, as they highlight spatial reference frames as important *representations*.

In summary, one could apply the CA model from Huttenlocher et al. (2000) as a computational level description for computing spatial term acceptabilities. The general claim would be: On the computational level, spatial prepositions are understood best, if they match their corresponding spatial categories. This explicates the goal of the overall task for the cognitive system (maximizing probability for finding LOs) – a discussion missing in Regier and Carlson (2001). Moreover, there is

*Spatial language
verification means
matching categorical
spatial prepositions
to fine-grained
locations.*

evidence supporting the general idea of language being optimal in categorizing real-world entities (Kemp & Regier, 2012; Khetarpal, Majid, & Regier, 2009; Regier & Xu, 2017).

6.2.3 *Explicating the Algorithmic and Representational Level: Reference Frames and Attentional Shifts*

The AVS-like models are already primarily specified on the algorithmic/representational level. This section aims to explicate existing representations and processes and to link them to relevant research. First, I review empirical findings based on predictions from the CA model – the model identified to be a suitable fit for the computational level. These findings highlight the important role of reference frames in spatial cognition.

More than 20 years ago, Logan (1995, p. 103) proposed a “theory [that] interprets [spatial] reference frames as mechanisms of attention, similar to spatial indices but with more computational power.” More precisely, the proposed “theory of voluntary, top-down control of visual spatial attention [...] explains how linguistic cues like ‘above,’ ‘below,’ ‘left,’ and ‘right’ are used to direct attention from one object to another” (Logan, 1995, p. 103). More recently, Gibson and colleagues extended this theory (Gibson & Sztybel, 2014, see Section 2.2.3). Reviewing several studies applying Posner’s spatial cueing paradigm to test their theory (see Section 2.1.1), Gibson and Sztybel (2014) suggest that the spatial reference frame (that needs to be imposed on the RO) controls shifts of attention in response to linguistic cues. This makes reference frames likely representations on which shifts of attention operate: Reference frames might structure space so that shifts of attention “know where to go”.

Shifts of visual attention operate on spatial reference frames.

Despite representing attentional shifts, AVS-like models lack a representation of a reference frame. This is why I suggest to add an explicit representation of a spatial reference frame. To this end, I review relevant research on spatial reference frames and sketch potential reference frame implementations in AVS-like models. Thereafter, in terms of processes, I highlight the important role of shifts of visual attention (operating on spatial reference frames). Throughout the section, I discuss the effects from the empirical study in Chapter 4 in light of the proposed explications of the algorithmic/representational-level.

Reference Frames

In research on linguistic and non-linguistic categorization of space, spatial reference frames – in the form of cardinal axes – were identified as important representations. For instance, Huttenlocher et al. (2004) investigated spatial categorization to test predictions from the CA model (a potential model of spatial language verification on the computational level). One prediction from the CA model is that stimuli reproduction

is biased towards the center of categories. Huttenlocher et al. (2004, see also Huttenlocher, Hedges, & Duncan, 1991) placed small dots in a circle and presented this to their participants for one second (only one dot visible per trial; see Figure 6.1).² Thereafter, participants had to re-create the location of the dot in an otherwise blank circle. Huttenlocher et al. (2004, 1991) found that the estimated locations were systematically biased towards the centers of the four quadrants of the circle (i.e., top right, bottom right, bottom left, top left; see Figure 6.1a). The closer a dot was placed to a cardinal axis, the more participants mis-placed it towards the center of the quadrant category. Thus, the cardinal axes (interpretable as the circle's intrinsic spatial reference frame) seem to serve as boundaries of the categories, while the diagonal axes – located at the centers of the quadrants – seem to be prototypes of the spatial categories.

Crucially, Huttenlocher et al. (2004) found that this spatial categorization did not change with a higher distribution of dots around the cardinal axes (a manipulation to make the diagonal axes better category boundaries; see Figure 6.1b): Even if participants were explicitly instructed to use the diagonal axes as category boundaries, their estimations remained biased to the center of the quadrants (but see Lipinski, Simmering, Johnson, & Spencer, 2010³). This suggests that the quadrants are strong a priori spatial categories (see also Crawford, Regier, & Huttenlocher, 2000). With respect to the linguistic vs. non-linguistic representation of space, Crawford et al. (2000, p. 209) state that their “findings suggest that while linguistic and non-linguistic spatial organization rely on a common underlying structure, that structure may play different roles in the two organizational systems”.

More to the point, the cardinal axes seem to serve as *boundaries* for non-linguistic spatial categorization but as *prototypical* examples for linguistic categorization (cf., Hayward & Tarr, 1995, summarized in Section 1.1.2). Munnich, Landau, and Doshier (2001) reason that linguistic success relies on clearly specified categories (cf. Section 6.2.2). The cardinal axes are salient geometric properties which might be a reason why linguistic encoding of space makes use of them (compare this with the use of salient intrinsic reference frames in spatial descriptions, e.g.,

Cardinal axes are important for linguistic and non-linguistic categorization of space.

2 See e.g., Kranjec et al. (2014) and van der Ham and Postma (2010) for studies with similar stimuli in the cognitive neuroscience tradition of categorical vs. coordinate spatial relations. See Franklin, Henkel, and Zangas (1995) for presenting related empirical evidence that locations relative to an egocentric reference frame (‘surrounding space’) follows a spatial structure similar to those found by Huttenlocher et al. (1991). Finally, see Feist and Gentner (2007) for biases towards centers of linguistic categories in a memory task with spatial relations.

3 While Huttenlocher et al. (2004) argue that the quadrants in a circle are robust spatial categories – immune to changes in distributions of dots –, Lipinski, Simmering, et al. (2010) consider longer time scales and show that different distributions can affect the use of these categories. (See also Lipinski, Spencer, & Samuelson, 2010a, for findings related to learning and Schutte & Spencer, 2009, for work that addresses developmental changes in spatial categorization.)

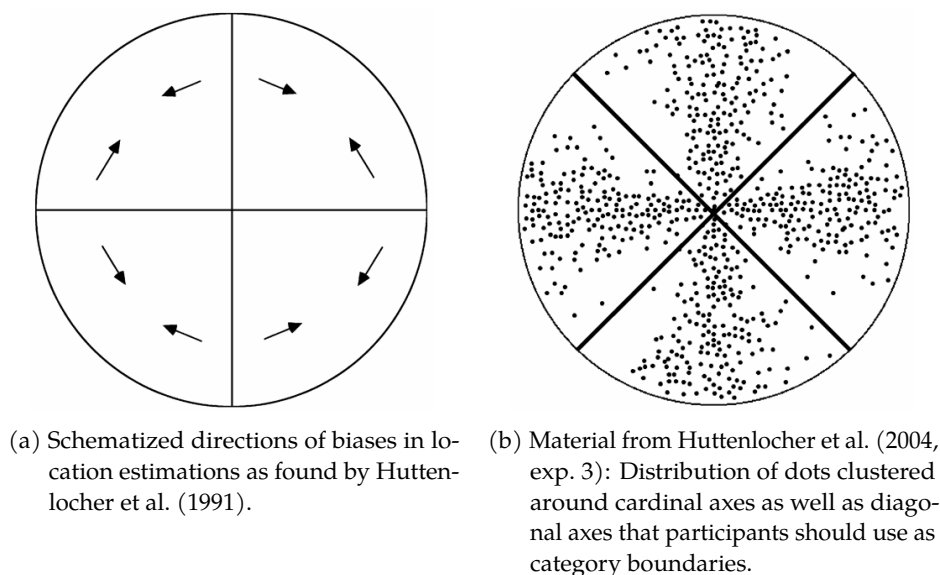


Figure 6.1: Visualization of stimuli and findings from location estimation studies by Huttenlocher et al. (2004, 1991); Lipinski, Simmering, et al. (2010). Image copyright: See Appendix E.

“The tree is in front of the bike”). On the other hand, their saliency also establishes the cardinal axes as very precise boundaries for spatial categories. Precise boundaries help in avoiding mis-categorizations of stimuli which could be the reason for a predominance of cardinal axes as boundaries in non-linguistic spatial categorization.

A different explanation for the seemingly conflicting data from Crawford et al. (2000) and Huttenlocher et al. (2004, cardinal axes as boundaries) vs. Hayward and Tarr (1995, cardinal axes as prototypes) is provided by Lipinski, Spencer, and Samuelson (2009, 2010b). Lipinski, Spencer, and Samuelson (2010b) show that while Crawford et al. (2000) and Hayward and Tarr (1995) used similar tasks in their investigations of (non-)linguistic representations of space, the tasks differed in an important aspect: the delay after which participants had to rate or recall a spatial relationship. Lipinski, Spencer, and Samuelson (2010b) argue that this makes it difficult to compare the empirical data from the tasks.

Accordingly, aiming for better-comparable linguistic rating tasks and non-linguistic spatial recall tasks, Lipinski, Spencer, and Samuelson (2010b) manipulated the delay after which participants had to respond – either immediately or after 10 seconds (see also Lipinski et al., 2009). Interestingly, they found that both linguistic and non-linguistic tasks were affected by the delay in the same direction. For the non-linguistic spatial recall task, people showed greater bias and greater variability in location estimation with an increasing delay. Correspondingly, for the longer delay, people gave lower acceptability ratings with a higher variability. Note that lower linguistic ratings correspond to a spatial

drift away from the cardinal axes which is the same pattern as found for the non-linguistic location estimations.

Lipinski, Spencer, and Samuelson (2010b) propose that linguistic and non-linguistic organization of space is driven by shared representational states (following Hayward & Tarr, 1995). Lipinski et al. (2009) support this claim by simulations of a computational model. Crucially, the representational states (contrasting static representations) depend on time – they are dynamic. This challenges the AVS and the CA model (which Lipinski et al., 2009, explicitly mention), which both lack a temporal component. The computational model proposed by Lipinski et al. (2009) is formulated in the ‘Dynamic Field Theory’ (DFT) framework. The DFT is a comprehensive framework to build neuronally plausible cognitive models. This is why I present the model from Lipinski et al. (2009) as a promising candidate for the implementational level in Section 6.2.4. In particular, the DFT model provides an elegant and neuronally plausible explanation of how reference frames might be represented and processed in the brain.

Another line of research asks whether the cross-linguistically different use of spatial reference frames affects non-linguistic cognition (e.g., Brown & Levinson, 1993; Levinson, 2003; Levinson, Kita, Haun, & Rasch, 2002; Li & Gleitman, 2002; Majid, Bowerman, Kita, Haun, & Levinson, 2004). For instance, Brown and Levinson (1993) tested speakers of Dutch and Tzeltal (a Mayan language spoken in Mexico). While Dutch (similar to English and German) uses a relative reference frame to describe scenes like “The glass is to the left of the plate”, Tzeltal uses an absolute reference frame: “The glass is north of the plate”.

The main idea behind the experimental paradigm used by Brown and Levinson (1993, and many others, see Levinson, 2003; Majid et al., 2004) is to let participants perform a spatial task (e.g., to remember a spatial relation like glass-left-of-plate) and thereafter to rotate participants by 180 degrees and let them do another spatial task (e.g., to re-create the remembered spatial relation). For Dutch speakers, the linguistic reference frame rotates with their bodies while for Tzeltal speakers the linguistic reference frame remains stable during the rotation. A solution to the task (e.g., remembering and re-creating a spatial relation) that relies on an absolute reference frame should result in the same absolute placement of the objects regardless of participants’ rotation – the rotation does not change the “north-object”. In contrast, after a 180 degrees rotation, a relative solution should place the objects with respect to the viewer – effectively reversing the absolute placement of the objects. Indeed, Brown and Levinson (1993) found that Tzeltal speakers predominantly solved the task with absolute solutions while Dutch speakers provided relative solutions. Based on this type of studies, Majid et al. (2004, p. 108) “argue that language can play a significant role in structuring, or restructuring, a domain as fundamental as spatial cognition”.

AVS-like models lack a temporal component.

Different languages use different reference frames, affecting their non-linguistic categorization of space.

ADDING REFERENCE FRAMES TO AVS-LIKE MODELS Taken together, spatial reference frames are important representations for spatial cognition and spatial language. In particular, reference frames are crucial for the interaction of linguistic and non-linguistic spatial organization. Despite this importance, AVS-like models lack an explicit reference frame representation (but see Schultheis & Carlson, 2018, who combine the AVS model with a model of reference frame selection).

AVS-like models should explicitly contain a spatial reference frame.

Remember that a reference frame is theorized to consist of four parameters (Logan & Sadler, 1996, see Section 2.2.1): origin, orientation, direction, and scale. Schultheis and Carlson (2017) provide evidence that in case of conflicting reference frames (e.g., absolute reference frame provided by gravity vs. relative reference frame provided by a reclining observer), humans do not select between complete reference frames but rather select single reference frame parameters based on the available information. Thus, it seems appropriate to integrate each reference parameter independently from the others into AVS-like models (instead of implementing complete reference frames).

However, the reference direction in AVS-like models (against which the direction of the vector sum is compared to in order to obtain an angular deviation) seems to encode both the orientation and the direction of the axis corresponding to the modeled preposition. While the orientation rotates the axes of a reference frame, the direction distinguishes above from below or left from right. Both, the orientation and the direction of the RO (e.g., Carlson-Radvansky & Logan, 1997) and the LO (e.g., Burigo et al., 2016; Burigo & Sacchi, 2013; Burigo & Schultheis, 2018) affect spatial language acceptability judgments. However, Burigo and Schultheis (2018) provide evidence that the orientation of the LO seems to be irrelevant for spatial language use while the direction of the LO has an effect. A future enhancement for AVS-like models should discuss these different contributions of orientation vs. direction to spatial language use by untangling the current intertwinement of both reference frame parameters in the models.

The empirical findings presented in Chapter 4 could inform the role of the two other reference frame parameters (origin and scale) in AVS-like models. The origin parameter could be interpreted as a reference point within the RO. Here, the rating data from the asymmetrical ROs suggest that the center-of-object might be a more appropriate reference point than the center-of-mass. In addition, the collected eye-tracking data could prove useful for identifying a preferred reference point in the RO (see also the discussion about spatial pooling in the next section). Finally, the test case related to relative distance revealed interesting details about the effect of the RO size. This could be an important source of information for the scale parameter of the reference frame. One idea to explicate the scale parameter is to encode it via the length of the vector sum or the reference direction – currently, these lengths do not matter for the model computations. After having discussed

reference frames, the next section considers a fundamental process that likely operates on spatial reference frames: shifts of visual attention.

Shifts of Visual Attention

Shifts of visual attention are important for processing spatial relations (see Section 2.2.3). Likely, attentional shifts operate on spatial reference frames (Gibson & Sztybel, 2014; Logan, 1995). Logan (1995, p. 103) interprets “reference frames [as being] similar to spatial indices”. Spatial indices are theorized to ground mental representations in the visual world (Pylyshyn, 2000, 2001; Spivey et al., 2004, see Section 2.1.1). Taken together, shifts of visual attention operating on reference frames are most likely grounding spatial language in the visual world.

Here, a correspondingly enhanced AVS-like model could further benefit from being combined with the CTVA model (Bundesen, 1998; Logan, 1996; Logan & Bundesen, 1996, see Section 2.1.1). The CTVA is rooted in visual attention research and thus could help to more explicitly link the notion of attention in AVS-like models to relevant research. On the other hand, the CTVA lacks an explicit account of the selection between objects (Logan, 1996, p. 623 and 635). Here, the AVS-like models are a useful addition, as their implementation of shifts of attention represents such selection. Moreover, “[i]n principle, CTVA should interface nicely with theories in which spatial indexing is an important process” (Logan, 1996, p. 615). Thus, I reckon integrating the CTVA with AVS-like models is a fruitful undertaking. This model combination would foster our understanding of how visual attention grounds spatial language in the visual world.

Particularly important for grounding spatial prepositions is the process of shifting visual attention. This importance is also reflected within the neuroscientific distinction of categorical vs. coordinate processing of spatial relations.⁴ Here, researchers found that small attentional windows enhance categorical processing while large attentional windows enhance coordinate processing (e.g., Laeng et al., 2011, see also Section 2.2.2). Crucially, with a small attentional window one cannot select the two objects of a spatial relation simultaneously. This suggests that categorical processing of spatial relations (e.g., processing spatial prepositions) depends on serial movements (i.e., shifts) of visual attention (see also Stocker & Laeng, 2017).

Potentially, visual attention actively controls whether humans process a given spatial relation categorically or coordinately: For linguistic tasks, attention might be more focused than for non-linguistic tasks, where attention might be more broadly distributed. This idea might also be a valuable perspective for research on linguistic vs. non-linguistic spatial categorization (e.g., Hayward & Tarr, 1995; Huttenlocher et al., 2004,

⁴ See also Kosslyn (2006, p. 1523): “In fact, I would not be surprised if the distinction between categorical and coordinate spatial relations provides insight into how linguistic categories bridge to perceptual representations.”

2000; Lipinski et al., 2009, see page 157f). This is because this type of research also distinguishes between coarse spatial categories and fine-grained encoded stimuli locations (e.g., in the Bayesian CA model, Huttenlocher et al., 2000, see Section 6.2.2).

Considering the main outcome of this Ph.D. project, it seems that the directionality of the attentional shift (from the RO to the LO vs. from the LO to the RO) does not matter for spatial language verification. However, the *orientation* of the attentional shift is important for spatial language verification. In AVS-like models, the orientation of the attentional shift is encoded in the orientation of the vector sum. The angular deviation of this orientation to the reference direction (as implemented in the models) is a major component affecting the model outcome. In terms of reference frame parameters, this orientation-focused mechanism favors the orientation parameter over the direction parameter (with respect to the importance for spatial language verification). This interpretation contrasts the findings from Burigo and Schultheis (2018) who found that the orientation of the LO seems to be irrelevant for spatial language use while the direction of the LO has an effect. Future research should specify how the reference frame parameters (with one reference frame per object) interact with the orientation and direction of the attentional shift (that moves from one object to the other).

THE IMPORTANCE OF THE VECTOR SUM As a final point in discussing the algorithmic/representational level, I want to highlight the role of the weighted vector sum. There are two important components of the weighted vector sum: averaging and directing.

In terms of averaging, the vector sum is related to studies in saccadic and perceptual localization (e.g., Cohen, Schnitzer, Gersch, Singh, & Kowler, 2007; Desanghere & Marotta, 2015; Melcher & Kowler, 1999; Nuthmann & Henderson, 2010; Vishwanath & Kowler, 2003). In these studies, researchers investigate how the visual system computes reference points on objects, e.g., preferred fixation landing positions when participants are instructed to “look at the object as a whole”. Across several different conditions, the center-of-mass of an object was identified as a preferred saccadic end point. The eye-gaze data presented in Chapter 4 support this evidence by showing that participants’ fixations were affected by the asymmetrical mass distribution of the L-shaped objects: The more object mass an area contained, the more fixations landed in this area (despite acceptability ratings not reflecting the asymmetrical mass distribution). However, Vishwanath and Kowler (2003) found that preferred reference points were also affected by the specific task given to the participants.⁵ Such task-dependency could explain why the gaze patterns for the less open C-shaped ROs from Chapter 4

⁵ Interestingly (with respect to the discussion of the importance of reference frames), Vishwanath and Kowler (2003, p. 1652) mention “a reference frame effect” as a possible explanation for biases in alignment judgments.

did not reflect the asymmetrical mass distribution. Rather, participants fixated the C-shaped ROs as if they were rectangular.

As a likely mechanism in the visual system for computing reference points, researchers proposed ‘spatial pooling’ (e.g., Melcher & Kowler, 1999; Vishwanath & Kowler, 2003). This mechanism averages information over a given spatial region. Given that the vector sum in AVS-like models is spatially restricted (to a single object only), the averaging component of the vector sum can be interpreted as spatial pooling. Interestingly, Cohen et al. (2007) showed that spatial pooling interacts with attention. Cohen et al. (2007) presented stimuli displays that contained both target and distractor objects. Crucially, the distractor objects were located within the same region as the target objects. However, for the experimental task, only target objects should be considered (while distractor objects should be ignored). The results from Cohen et al. (2007) suggest that instead of pooling across an entire spatial region, an attentional filter enables the visual system to selectively average information from target objects only. As a consequence, Cohen et al. (2007) suggest a spatial pooling mechanism that is weighted by attention. This reminds of the attentional weighting of the vector sum in AVS-like models. Future research should combine these two lines of research. AVS-like models would profit from being explicitly linked to mechanisms of overt attention (i.e., saccades) that specify how the visual system “selects” objects.

The attentional vector sum is similar to attentionally weighted spatial pooling.

The second component of the vector sum is its direction which encodes the direction of the attentional shift. Apart from this interpretation, the use of a *vector* sum aligns well with existing vector-based approaches of spatial cognition in general and the semantics of spatial prepositions in particular (e.g., O’Keefe, 1996, 2003; Zwartz, 1997, 2017; Zwartz & Gärdenfors, 2016). In particular, the ‘vector grammar’ theory put forward by O’Keefe (1996, 2003) could be a promising candidate to extend AVS-like models (i) with further spatial prepositions and (ii) with connections to neuroscientific literature. Connections to neuroscientific literature can be made because John O’Keefe is known for discovering place cells in the hippocampus, a discovery for which he was awarded the 2014 Nobel Prize in Physiology and Medicine (Kiehn & Forssberg, 2014). O’Keefe shares this award with May-Britt Moser and Edvard I. Moser, who discovered grid cells. Together, place and grid cells are thought to constitute the brain’s spatial representation system: these cells might be the neuronal substrate of a ‘cognitive map’ (Moser, Kropff, & Moser, 2008; O’Keefe & Nadel, 1978, 1979; Tolman, 1948).

Developing a computational model of this hypothesized cognitive map, O’Keefe (1990, p. 310) also writes about a “grand vector sum of a large group of place cell fields”. Furthermore, it has been proposed that in such a cognitive map, navigation is based on vectors – in particular, based on the role of the grid cells (e.g., Bush, Barry, Manson, & Burgess,

2015). Recently, this claim was supported by simulations of artificial neural networks that were built to enable navigation for artificial agents (Banino et al., 2018; Cueva & Wei, 2018; Savelli & Knierim, 2018). Without explicitly modeling grid cells, researchers found that these artificial neural networks developed grid-like representations resembling the firing pattern of natural grid cells. Taken together, this supports the general importance of vector-based representations for human spatial skills.

6.2.4 *Extending the Implementational Level: Grounding Spatial Language*

Marr's implementational level asks how the proposed model representations and mechanisms could be implemented in the brain. Apart from the connections of the vector (sum) representation to the brain's navigational system as outlined above, there exist further possible ties to specify AVS-like models on the implementational level. In particular, I want to highlight the potential to explicitly link AVS-like models to computational models formulated in the 'Dynamic Field Theory' (DFT) framework. The DFT framework is based on principles of neuronal dynamics (for introductions to the DFT see Schöner, 2008; Schöner, Spencer, & the DFT Research Group, 2016). Thus, models built within the DFT framework can be considered to be consistent with how (parts of) the brain work.⁶

In particular, the DFT model of spatial recall proposed by Lipinski et al. (2009) reveals interesting insights into how spatial reference frames might be represented in the brain. Lipinski et al. (2009, see also Lipinski, Spencer, & Samuelson, 2010b) were motivated by the seemingly contrasting data from Hayward and Tarr (1995) – cardinal axes are *prototypes* – vs. Crawford et al. (2000) – cardinal axes are category *boundaries* (see also Huttenlocher et al., 2004, and Section 6.2.3). In the DFT model from Lipinski et al. (2009), spatial working memory encodes "perceived reference axes" (p. 114) – i.e., the cardinal axes. These encoded reference axes affect the subsequent encoding of the LO location (that human participants need to estimate after a short time interval). More precisely, the neuronally plausible mechanisms from the DFT framework provide the following interaction pattern: Two activation peaks (i.e., neuronal representations of the reference axis and

⁶ A different framework with overlapping goals to the DFT framework is the 'Neural Engineering Framework' (or 'Semantic Pointer Architecture') proposed by Chris Eliasmith and colleagues (Eliasmith, 2015; Eliasmith & Anderson, 2004). Within this framework, Eliasmith et al. (2012) created a model of the brain (called Spaun) with 2.5 million simulated neurons. Using a robotic arm, this model can perform eight behavioral tasks (like copying digits, remembering lists, or answering simple questions). However, there has not been much work trying to represent language in this framework. Given that Eliasmith and Anderson (2004, Chapter 2.5) explicitly model the vector sum coding proposed by Georgopoulos et al. (1986, and referred to by Regier & Carlson, 2001, to motivate the vector sum in the AVS model), modeling spatial prepositions could be a valuable addition to the NEF "universe".

the LO location) either repel each other or, if close enough, attract each other. On the one hand, the repulsion explains why human location estimates are biased away⁷ from the cardinal axes. On the other hand, the attraction explains increased memory accuracy found for locations sufficiently close to cardinal axes (e.g., Hayward & Tarr, 1995). In summary, Lipinski et al. (2009) present a neuronally plausible model in which the same representation of a reference frame accommodates two empirical findings that were previously thought to be contradictory. In other words: “Spatial Language and Spatial Memory Use the Same Perceptual Reference Frames” (part of the title of Lipinski et al., 2009).

Lipinski et al. (2009) explicitly discuss the relationship of their DFT model to the AVS model. While Lipinski et al. (2009, p. 129) state that “there is conceptual overlap” between the two models, they also argue that the two models differ in how they compute acceptability ratings. I think that both models could profit from being analyzed using Marr’s three-level framework. In particular, I propose to interpret DFT models primarily as model on the implementational level and AVS-like models primarily as models on the algorithmic/representational level.

Such an analysis could help to untangle to what extent the two models actually differ or whether they specify similar mechanisms and representations on different levels of details. To the point, I believe that the DFT model from Lipinski et al. (2009) provides an elegant and neuronally plausible explanation of how reference frames might be represented and processed in the brain. AVS-like models can rely on these mechanisms by specifying computations using reference frames as more abstract representations. Such an approach would avoid that two models compete that are vastly different in terms of mathematical formulations. Rather, this analysis has the potential to consistently explain the phenomenon on different levels of description and thus to provide a more comprehensive account of spatial language processing (see also Newell, 1973).

This should not prevent future research from comparing different models of the process. Rather, I believe that such a comparison should specify explicit links between representations and algorithms of models on different levels instead of aiming for one single model to be “better” than the other model. Within this spirit, Lipinski et al. (2012, p. 1508) present a DFT model that is “highly compatible with the AVS model” while it offers neuronally plausible formulations for representations and mechanisms of AVS-like models (e.g., the vector sum; see Richter, 2018; Richter et al., 2014, 2016, 2017, for refined versions of this DFT model). Future research should further pursue this path. Based on the outcome of this Ph.D. project, it would be particularly interesting to

*In DFT models,
spatial reference
frames emerge in a
neuronally plausible
manner.*

⁷ In contrast to Crawford et al. (2000) and Huttenlocher et al. (2004, 1991) who propose a bias *towards* categorical prototypes at the diagonal axes, Lipinski et al. (2009) and Lipinski, Spencer, and Samuelson (2010b) argue for a bias *away* from the cardinal axes.

examine whether the directionality of the attentional shift also does not affect model performance for DFT models.

6.3 SUMMARY OF IDEAS FOR FUTURE MODEL ENHANCEMENTS

Throughout this chapter, I proposed to refine AVS-like models in several aspects. This section summarizes the proposed refinements. In addition, I suggest further model extensions.

In terms of representations in AVS-like models, I highlighted the absence of spatial reference frames. Reference frames are important representations in spatial cognition (e.g., Brown & Levinson, 1993; Crawford et al., 2000; Hayward & Tarr, 1995; Huttenlocher et al., 2004; Levinson, 2003; Lipinski et al., 2009; Majid et al., 2004). This is why I argue that AVS-like models should be extended with explicit representations of reference frames – or, better, with independent parameters of reference frames (Schultheis & Carlson, 2017). More precisely, the reference direction of AVS-like models should distinguish between the reference frame parameters orientation and direction. The implementation of the reference frame parameters origin and scale could be informed by the empirical findings presented in Chapter 4.

Following work by Logan (1995) and Gibson and Sztybel (2014), reference frames are likely representations on which shifts of visual attention operate. Future extensions to AVS-like models should consider to what extent the direction and orientation of reference frames interact with the direction and orientation of attentional shifts. In terms of visual attention, the CTVA model proposed by Logan (1996) seems to be a promising candidate to more tightly link the notion of attention in AVS-like models to visual attention research. One role of visual attention seems to be to control whether humans process spatial relations categorically or coordinately. Evidence from cognitive neuroscience suggests that for categorical processing, shifts of attention are necessary (e.g., Laeng et al., 2011).

The mechanism that implements attentional shifts in AVS-like models is the attentionally weighted vector sum. In terms of averaging, this mechanism is compatible with findings from saccadic and perceptual localization (e.g., Cohen et al., 2007). Future modeling should strengthen this connection to more explicitly link attentional shifts to saccadic eye movements. In terms of providing a direction to space, the vector sum as a vector-based representation is consistent with other vector-based approaches to spatial cognition (e.g., O'Keefe, 2003). The close connection of these approaches to what is thought to be the brain's spatial representation system (e.g., Moser et al., 2008; O'Keefe & Nadel, 1978) further supports the use of vectors. In addition, taking advantage of this connection could help to identify possible neuronal implementations of mechanisms from AVS-like models.

At this point, computational models specified in the DFT framework (e.g., Lipinski et al., 2012, 2009) could become useful, as they are based on neuronal dynamics. Due to their inherent dynamic properties, these models highlight the dynamic nature of cognition. Hence, they could serve as a useful information source for integrating a temporal component into AVS-like models. This is in particular important as shifts of attention are inherently temporal, too. A temporal component in AVS-like models would allow to investigate the temporal dynamics of attentional shifts in the context of spatial language verification. Such temporal information could be more directly linked to studies from the psycholinguistic visual world paradigm. Possible research questions could be: At what point in time does a spatial preposition trigger a shift of attention or how long does a linguistically-triggered shift of attention take?

The main outcome of this Ph.D. project is that the directionality of the shift does not matter for spatial language verification. Based on task demands, it might be that the directionality of the attentional shift can be flexibly adjusted. Thus, a fruitful next step for investigating shifts of attention during spatial language verification would be to create a model that flexibly allows for shifts in both directionalities (see also Burigo & Knoeferle, 2015, who observed both directionalities). This point of view fits with findings that visual attention affects categorical vs. coordinate processing of spatial relations. Finally, an analysis of the task of spatial language verification (affecting deployment of visual attention) on Marr's computational level certainly provides interesting insights into what the cognitive system computes – and why. Here, the CA model by Huttenlocher et al. (2000) seems to be a promising starting point.

6.4 CONCLUSION: DOES DIRECTIONALITY OF ATTENTION MATTER?

This Ph.D. project was primarily motivated by conflicting evidence regarding the role of the directionality of attentional shifts for spatial language verification. Shifts of attention are considered to be important for the processing of spatial relations (e.g., Franconeri et al., 2012; Logan & Sadler, 1996). Spatial language distinguishes the two objects of a spatial relation from each other as a reference object and a located object. The influential work by Gordon Logan claimed that attention should shift from the reference object to the located object (Logan, 1995; Logan & Sadler, 1996; Logan & Zbrodoff, 1999). Accordingly, Regier and Carlson (2001) developed the Attentional Vector Sum (AVS) model that (implicitly) realizes a shift of attention from the reference object to the located object. However, recent evidence challenges the directionality of this attentional shift. In contrast, attention might shift in the same order as the spatial sentence unfolds – i.e., from the located object to the reference object (Burigo & Knoeferle, 2015; Roth & Franconeri, 2012).

This is why I developed the reversed AVS (rAVS) model, a modification of the AVS model in which attention shifts from the located object to the reference object.

After assessing both shift implementations with an empirical study and model simulations, the main conclusion is that both directionalities of attention accommodate the empirical data equally well. On the one hand, this challenges the claim by Logan (1995), Logan and Sadler (1996), and Logan and Zbrodoff (1999). On the other hand, it provides support for the mechanisms shared by all models (in particular: spatial averaging and vector-based approach). Furthermore, the empirical and computational studies revealed two novel effects on spatial language verification related to relative distance and asymmetrical reference objects. Although these two effects are not directly related to the question about the directionality of the attentional shift, they motivated further model refinements; the resulting models performed substantially better on the empirical data than their predecessors.

In discussing the results, I provided several ideas for further model extensions. In particular, I sketched an analysis using David Marr's seminal three-level framework (Marr, 1982). Doing so revealed several promising connections to related research that could be exploited for future model extensions. The main motivation for all these model extensions is to create a more comprehensive model of spatial language verification. Such a model allows cognitive scientists to more closely analyze the role of shifts of visual attention for spatial language verification – as part of a broader research agenda that asks how humans ground their language to the visual world.

Part IV

APPENDIX



LIST OF ABBREVIATIONS

ABC	Approximate Bayesian Computation (Palestro et al., 2018), page 147
AVS-BB	AVS bounding box (considering the center-of-object instead of the center-of-mass), page 116
BB	bounding box of an RO, smallest rectangle containing all points of the RO, page 96
CI	credible interval, page 90
CoO	center-of-object; center of the bounding box of an RO, page 96
DFT	Dynamic Field Theory (Schöner et al., 2016), page 165
GOF	goodness-of-fit, page 41
KL divergence	Kullback-Leibler divergence, page 141
LO	located object (e.g., the bike in “The bike is in front of the house”), page 3
LOO	leave-one-out cross-validation method, a goodness-of-fit measure for statistical models adjusted for overfitting (Vehtari et al., 2017), page 90
MCMC	Markov Chain Monte Carlo, page 42
MFA	Model Flexibility Analysis (Veksler et al., 2015), page 123
NHST	Null Hypothesis Significance Testing, page 88
nRMSE	normalized Root Mean Square Error, page 41
PBCM	Parametric Bootstrap Cross-fitting Method (Wagenmakers et al., 2004), page 132
PSP	Parameter Space Partitioning (Pitt et al., 2006), page 81
rAVS-CoO	rAVS center-of-object (considering the center-of-object instead of the center-of-mass), page 117
RC-LS	parameter set reported in Regier and Carlson (2001) as best fit to data from Logan and Sadler (1996, exp. 2, <i>above</i>), see Table 3.1, page 48

RO	reference object (e.g., the house in “The bike is in front of the house”), page 3
SHO	simple hold-out, page 41

EMPIRICAL STUDY

The following German text is the introductory text every participant read before the experiment began:

Hallo und vielen Dank, dass Du an meiner Studie teilnimmst!

In dieser Studie werden Dir Bilder zusammen mit entsprechenden Sätzen gezeigt und Du musst bewerten, wie gut ein Satz das jeweilige Bild beschreibt. Zur Bewertung gibt es eine Skala von 1 bis 9, wobei 1 bedeutet, dass der Satz das Bild überhaupt nicht beschreibt und 9 bedeutet, dass der Satz das Bild perfekt beschreibt. Um Deine Bewertung abzustufen, darfst und sollst Du gerne auch die Zahlen zwischen 1 und 9 nutzen. Beachte bitte, dass es keine „richtige“ oder „falsche“ Bewertung gibt. Wähle die Bewertung, die Deiner Meinung nach am Besten dazu passt, wie gut der Satz das Bild beschreibt.

Zum Ablauf: Es wird jeweils ein Satz auf dem Bildschirm angezeigt, den Du bitte aufmerksam liest. Nachdem Du den Satz gelesen und verstanden hast, drücke die Leertaste. Dann erscheint ein Bild, das Du mit den Ziffern 1 bis 9 daraufhin bewerten sollst, wie gut der vorher gelesene Satz dieses Bild beschreibt. Benutze dafür bitte die Ziffern auf der Tastatur über den Buchstaben. Wenn Du eine Ziffer gedrückt hast, erscheint der nächste Satz. Benutze bitte während des ganzen Experiments nur eine Hand. Bevor es richtig losgeht gibt es einige Durchgänge zum Ausprobieren.

Wenn Du noch Fragen zur Bewertung oder zum Ablauf hast, dann darfst Du diese jetzt gerne stellen.

Ansonsten folgt eine kurze Erläuterung zur Kalibrierung der Augenbewegungskamera. Dazu drücke bitte die Leertaste.

[nächster Bildschirm]

Während Du die Bilder bewertest, werden Deine Augenbewegungen aufgenommen. Dazu ist es wichtig, dass Du während des ganzen Experiments Deinen Kopf so still wie möglich hältst. Deine Augen darfst Du bewegen, Deinen Kopf bitte möglichst wenig. Um die Kamera zu kalibrieren, erscheinen gleich nacheinander 10 kleine Kreise an verschiedenen Stellen des Bildschirms (zu jedem Kreis gibt es auch einen Ton). Deine Aufgabe ist es, auf jeden Kreis zu schau-

en, indem Du Deine Augen dorthin bewegst (den Kopf aber bitte still halten). Diese Prozedur wird mindestens noch einmal wiederholt (also nochmals 10 Kreise), es könnte aber auch nötig sein, dass die Prozedur öfter wiederholt wird.

Bevor die Kamera kalibriert wird, muss die Kamera noch eingestellt werden.

Wenn während des Experiments ein kleiner Kreis in der Mitte erscheint, hast Du die Möglichkeit eine Pause zu machen. Du musst keine Pause machen, darfst das aber gerne tun. Ich werde Dich an den Stellen fragen, ob Du eine Pause machen möchtest.

Falls Du jetzt das Gefühl hast, irgendwie unbequem zu sitzen, gibt es noch die Möglichkeit etwas zu verstellen. Ab jetzt dauert das Experiment ca. 50 Minuten.

Falls Du sonst noch irgendwelche Fragen hast, darfst Du diese gerne jetzt stellen.

Vielen Dank fürs Teilnehmen und viel Spaß.

MODEL FLEXIBILITY ANALYSIS

Table C.1: Results of the Model Flexibility Analysis (MFA) computed using 50^4 model predictions and parameter ranges in Equations 3.15–3.17 (page 42) and 5.1–5.2 (page 118). The lower ϕ , the less flexible is the model. ϕ_1 follows suggestion by Veksler et al. (2015) to split each dimension of the data space into $\sqrt[3]{50^4}$ cells. ϕ_2 uses the range of the rating scale as domain specific number of cells per data space dimension. ϕ_{n_2} normalizes ϕ_2 by dividing with maximal possible $\phi_{2\max}$. ϕ_1 and ϕ_{n_2} are plotted in Figures 5.3a and 5.3b (page 129).

	all stimuli	rectangular ROs	asymmetrical ROs	Regier and Carlson (2001) stimuli
# of dimensions (n)	448	224	224	337
# of cells per dimension ($\sqrt[3]{50^4}$)	1.03555	1.07236	1.07236	1.04753
AVS ϕ_1	2.29136×10^{-3}	3.3088×10^{-4}	7.91520×10^{-4}	4.19520×10^{-4}
rAVS _{w-comb} ϕ_1	1.61680×10^{-3}	4.7568×10^{-4}	1.96000×10^{-4}	2.92480×10^{-4}
AVS-BB ϕ_1	4.33120×10^{-4}	3.3088×10^{-4}	2.00000×10^{-5}	3.39520×10^{-4}
rAVS-CoO ϕ_1	6.82080×10^{-4}	4.7568×10^{-4}	1.13600×10^{-5}	2.56480×10^{-4}
# of cells per dimension (range of rating scale)	9	9	9	10
$\phi_{2\max}$	1.97000×10^{-421}	1.11000×10^{-207}	1.11000×10^{-207}	6.25000×10^{-331}
AVS ϕ_2	$2.834010 \times 10^{-422}$	$7.144000 \times 10^{-209}$	$1.112350 \times 10^{-208}$	$3.221390 \times 10^{-332}$
AVS ϕ_{n_2}	1.438584×10^{-1}	6.436036×10^{-2}	1.002117×10^{-1}	5.154224×10^{-2}
rAVS _{w-comb} ϕ_2	$2.572760 \times 10^{-422}$	$8.904460 \times 10^{-209}$	$6.874580 \times 10^{-209}$	$2.711600 \times 10^{-332}$
rAVS _{w-comb} ϕ_{n_2}	1.305970×10^{-1}	8.022036×10^{-2}	6.193315×10^{-2}	4.338560×10^{-2}
AVS-BB ϕ_2	$1.501890 \times 10^{-422}$	$7.144000 \times 10^{-209}$	$6.887150 \times 10^{-210}$	$3.049380 \times 10^{-332}$
AVS-BB ϕ_{n_2}	7.623807×10^{-2}	6.436036×10^{-2}	6.204640×10^{-3}	4.879000×10^{-2}
rAVS-CoO ϕ_2	$1.894780 \times 10^{-422}$	$8.904460 \times 10^{-209}$	$5.396280 \times 10^{-210}$	$2.590890 \times 10^{-332}$
rAVS-CoO ϕ_{n_2}	9.618173×10^{-2}	8.022036×10^{-2}	4.861514×10^{-3}	4.152067×10^{-2}

Table C.2: Results of the Model Flexibility Analysis (MFA) computed using 50^4 model predictions and smaller parameter ranges in Equations 5.4–5.6 (page 127). The lower ϕ , the less flexible is the model. ϕ_1 follows suggestion by Veksler et al. (2015) to split each dimension of the data space into $\sqrt[3]{50^4}$ cells. ϕ_2 uses the range of the rating scale as domain specific number of cells per data space dimension. ϕ_{n_2} normalizes ϕ_2 by dividing with maximal possible ϕ_{2max} . ϕ_1 and ϕ_{n_2} are plotted in Figures 5.3c and 5.3d (page 129).

	all stimuli	rectangular ROs	asymmetrical ROs	Regier and Carlson (2001) stimuli
# of dimensions (n)	448	224	224	337
# of cells per dimension ($\sqrt[3]{50^4}$)	1.03555	1.07236	1.07236	1.04753
<hr/>				
AVS ϕ_1	3.07088×10^{-3}	3.7568×10^{-4}	1.04784×10^{-3}	1.27984×10^{-3}
rAVS _{w-comb} ϕ_1	2.29648×10^{-3}	5.9792×10^{-4}	2.06880×10^{-4}	1.06016×10^{-3}
AVS-BB ϕ_1	4.77920×10^{-4}	3.7568×10^{-4}	2.06400×10^{-5}	1.00496×10^{-3}
rAVS-CoO ϕ_1	8.76640×10^{-4}	5.9792×10^{-4}	1.12000×10^{-5}	9.12640×10^{-4}
# of cells per dimension (range of rating scale)	9	9	9	10
<hr/>				
ϕ_{2max}	1.97000×10^{-421}	1.11000×10^{-207}	1.11000×10^{-207}	6.25000×10^{-331}
AVS ϕ_2	6.51392×10^{-422}	1.09743×10^{-208}	2.33600×10^{-208}	8.84227×10^{-332}
AVS ϕ_{n_2}	0.3308271	0.09886757	0.2104505	0.141502
rAVS _{w-comb} ϕ_2	7.50826×10^{-422}	2.27392×10^{-208}	1.46568×10^{-208}	8.19759×10^{-332}
rAVS _{w-comb} ϕ_{n_2}	0.3809524	0.2048577	0.1320432	0.1312253
AVS-BB ϕ_2	2.37832×10^{-422}	1.09743×10^{-208}	8.82670×10^{-210}	8.44176×10^{-332}
AVS-BB ϕ_{n_2}	0.1203008	0.09886757	0.007951982	0.1350682
rAVS-CoO ϕ_2	5.10068×10^{-422}	2.27392×10^{-208}	7.15298×10^{-210}	7.93962×10^{-332}
rAVS-CoO ϕ_{n_2}	0.2581454	0.2048577	0.006444126	0.1272727

PRESENTED ON
MAY 10, 2019

THESIS 1: MODELING THE CONTRIBUTION OF VISUAL ATTENTION TO SPATIAL LANGUAGE VERIFICATION

This Ph.D. project focused on the contribution of visual attention to spatial language verification. Consider the sentence “The circle is above the rectangle” and imagine a corresponding spatial scene. It has been argued that humans shift their visual attention from the rectangle to the circle during the processing of such a sentence. However, recent empirical evidence suggests that attention might shift in the reversed direction – from the circle to the rectangle. Thus, this Ph.D. project addresses the following question: Does the direction of the attentional shift matter? Using computational cognitive modeling as well as empirical research, this Ph.D. project concludes that both directionalities of attention accommodate the existing empirical data equally well.

THESIS 2: EARLY WORD LEARNING: CHILDREN RELY ON STRUCTURAL PROPERTIES OF LEXICAL-SEMANTIC NETWORKS

In semantic processing tasks (e.g., free association or semantic categorization tasks), adults respond faster to earlier-acquired words than to later-acquired words (e.g., Brysbaert, Van Wijnendaele, & De Deyne, 2000). Using a natural reading task, Dirix and Duyck (2017) recently provided evidence for shorter fixations while reading earlier-learned words compared to later-learned words. Importantly, the age-of-acquisition of a word predicts people’s latencies independent of and “above other important (correlated) lexical variables, such as word frequency and length” (Dirix & Duyck, 2017, p. 1915).

Thus, the order in which children learn words lays an important foundation for adult language processing. This relation of early word learning with adult performance in semantic processing tasks can be explained by assuming *semantic networks* (e.g., Steyvers & Tenenbaum, 2005). In these semantic networks, earlier-learned words are better connected than later-learned words. Thus, a semantic search (e.g., via spreading activation, Collins & Loftus, 1975) starting from an earlier-learned word finishes faster than a search starting from a later-learned word.

This semantic network approach is inspired by network theory (or graph theory) – a mathematical method applied in a multitude of scientific disciplines (e.g., physics, theoretical computer science, biology, sociology, linguistics) in order to analyze complex systems. In its simplest form, a network consists of nodes (e.g., words) and edges connecting the nodes (e.g., associations between words). The degree of a node is the number of edges that connect it with other nodes.

Watts and Strogatz (1998) identified interesting network properties found in many natural networks (from neural networks to collaboration networks of film actors to the world-wide-web, Adamic, 1999; Watts & Strogatz, 1998). These networks have small average shortest path lengths (a global property measuring the distance between any two nodes) and are highly clustered (a local property: well connected neighborhoods of nodes). Based on sociological research by Milgram (1967), Watts and Strogatz (1998) call networks with these specific properties *small-world networks*. In addition to being a small-world network, the world-wide-web also possesses another interesting property: It is *scale-free*, i.e., it has a degree distribution that follows a power law (Albert, Jeong, & Barabási, 1999; Barabási & Albert, 1999). Intuitively, a scale-free network has a small amount of “hub-nodes” (connected to many other nodes) and many nodes that have relatively few connections.

Based on these findings from network theory, Steyvers and Tenenbaum (2005) could show that semantic networks created from linguistic data (word associations, WordNet, and Roget’s Thesaurus) have a small-world structure and are scale-free. Building on an algorithm that creates scale-free networks proposed by Barabási and Albert (1999), Steyvers and Tenenbaum (2005) suggest that a similar mechanism guides early word learning. This “model of semantic growth” is called *preferential attachment*. Preferential attachment assumes that words are more likely to be learned by children, if these words link to already well-connected words in the semantic network of known words (compared to learning words that connect to known words with less connections). That is, known words that have many connections to other known words are more likely to receive even more connections from newly learned words (compared to known words with less connections). Preferential attachment is also known as “the rich gets richer” and it generates scale-free networks (Barabási & Albert, 1999). A scale-free network would explain the earlier described effect of age-of-acquisition in adult semantic performance: Earlier-learned words (“hub-nodes”) are semantically better connected than later-learned words.

While not denying that such a network structure explains age-of-acquisition effects, Hills, Maouene, Maouene, Sheya, and Smith (2009) argue that there might be more suitable mechanisms for how semantic networks could grow (i.e., in which order children learn words). More specifically, they propose two new mechanisms, called *preferential acquisition* and the *lure-of-the-associates*. Both of these new mechanisms consider the learning environment of the children (formalized as an external semantic network). Preferential acquisition assumes that words are more likely to enter the lexicon, if they are well connected within the learning environment (i.e., they are linked to many words the children do not yet know). The lure-of-the-associates model lies between preferential acquisition and preferential attachment: It assumes that what matters are the number of connections from all known words to the words in the learning environment.

Analyzing the growth of networks of nouns (typically known by 16–30 months-old children), Hills et al. (2009, see also Amatuni & Bergelson, 2017) found that preferential acquisition and the lure-of-the-associates better described early word learning compared to preferential attachment (but see Sailor, 2013). These findings were corroborated by Hills, Maouene, Riordan, and Smith (2010), who showed that the contextual diversity of a to-be-learned-word (i.e., in how many different contexts the word is used) predicts the order of early word learning.

Based on the above reviewed studies, I argue that preferential attachment – despite its universal application throughout scientific disciplines (Barabási & Albert, 1999) – is not an appropriate model for describing early word learning. I do this for the following two reasons: First, the statistical structure of child-directed speech is important for early word learning (e.g., Romberg & Saffran, 2010). However, preferential attachment ignores the statistical structure of child-directed speech by only considering the internal semantic network (words that are already known by the child). Second, although preferential attachment as originally proposed by Barabási and Albert (1999) was considered a universal growth principle for natural (scale-free) networks, Keller (2005) convincingly argues that the scale-free property is less special than thought. Furthermore, there exist many different growth models that generate scale-free networks (Keller, 2005). Taken together, the mechanisms of preferential acquisition or the lure-of-the-associates seem to better describe early word learning than preferential attachment.

REFERENCES (THESIS 2: SEMANTIC GROWTH)

- Adamic, L. A. (1999). The small world web. In S. Abiteboul & A.-M. Vercoustre (Eds.), *Research and advanced technology for digital libraries. ECDL 1999*. (pp. 443–452). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-48155-9_27
- Albert, R., Jeong, H., & Barabási, A.-L. (1999). Diameter of the world-wide web. *Nature*, *401*(6749), 130–131. <https://doi.org/10.1038/43601>
- Amatuni, A., & Bergelson, E. (2017). Semantic networks generated from early linguistic input. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 1538–1543). Austin, TX, USA: Cognitive Science Society.
- Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. <https://doi.org/10.1126/science.286.5439.509>
- Brysaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, *104*(2), 215–226. [https://doi.org/10.1016/S0001-6918\(00\)00021-4](https://doi.org/10.1016/S0001-6918(00)00021-4)
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. <https://doi.org/10.1037/0033-295X.82.6.407>
- Dirix, N., & Duyck, W. (2017). An eye movement corpus study of the age-of-acquisition effect. *Psychonomic Bulletin & Review*, *24*(6), 1915–1921. <https://doi.org/10.3758/s13423-017-1233-8>
- Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, *63*(3), 259–273. <https://doi.org/10.1016/j.jml.2010.06.002>
- Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science*, *20*(6), 729–739. <https://doi.org/10.1111/j.1467-9280.2009.02365.x>
- Keller, E. F. (2005). Revisiting “scale-free” networks. *BioEssays*, *27*(10), 1060–1068. <https://doi.org/10.1002/bies.20294>
- Milgram, S. (1967). The small world problem. *Psychology Today*, *1*(1), 61–67.
- Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914. <https://doi.org/10.1002/wcs.78>
- Sailor, K. M. (2013). Is vocabulary growth influenced by the relations among words in a language learner’s vocabulary? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1657–1662. <https://doi.org/10.1037/a0032993>
- Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78. https://doi.org/10.1207/s15516709cog2901_3
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, *393*(6684), 440–442. <https://doi.org/10.1038/30918>

THESIS 3: LANGUAGE INFLUENCES HOW WE PERCEIVE COLORS: EVIDENCE FOR PROBABILISTIC INFERENCE

While English has one color term for *blue*, Russian has two distinct color terms: *goluboy* (light blue) and *siniy* (dark blue). Crucially, Russian does not have a single term that describes both shades of blue at the same time. Given this cross-linguistic difference, do native Russian speakers perceive blue colors differently than English native speakers?

According to Winawer et al. (2007), the answer to this question is “yes”. In a speeded color discrimination task, Winawer et al. (2007) presented three blue-colored squares in a triangle configuration (one square at the top, two squares at the bottom). The task of the participants (native Russian speakers and native English speakers) was to determine which of the two bottom squares was colored in the same way as the top square. One of the two bottom squares always showed the same color as the top color, i.e., the task had an objective solution. The second, distracting color was either from the same Russian category as the target color (within-category trial) or from the other color category (across-category trial).

Compared to within-category trials, Winawer et al. (2007) found that Russian speakers responded faster in across-category trials, i.e., if the distracting color was from the different color category. Crucially, English native speakers did not show this effect of categorical perception. While a spatial interference task (remembering spatial patterns) did not change the pattern of results, a verbal interference task (rehearsing digits) disrupted the categorical perception of Russian speakers. Taken together, Winawer et al. (2007) interpret their findings as an influence of linguistic color categorization on color perception: Russian speakers perceive blue differently than English speakers.

The study by Winawer et al. (2007) is part of a long and controversial debate on the influence of language on perception and thought, also known as the *Sapir-Whorf hypothesis* or the *linguistic relativity hypothesis* (e.g., Kay & Kempton, 1984, for reviews see Boroditsky, 2006; Wolff & Holmes, 2011). The domain of color perception is one of the mayor testbeds for this hypothesis (Regier & Kay, 2009; Witzel, 2018): Given the diverse color naming patterns in different languages, do they lead to different color perception in the respective linguistic communities? If yes, to what extent?

With regard to color perception, Regier and Kay (2009) argue that the Sapir-Whorf hypothesis conflates the following two, more fine-grained questions: “1. Do color terms affect color perception? 2. Are color categories determined by largely arbitrary linguistic convention?” (p. 439). According to Regier and Kay (2009), a universalist would answer “no” to both questions and a relativist would answer “yes” to both questions. However, evidence seems to suggest a more diverse pattern of answers, supporting both the universalist and the relativist stances at the same time (Regier & Kay, 2009). With respect to the second question, Zaslavsky, Kemp, Tishby, and Regier (2019) note that “[l]anguages vary widely in the ways they partition colors into categories” (p. 208). However, Zaslavsky et al. (2019) argue that these linguistic variations are not based on arbitrary linguistic conventions. Using an information-theoretic analysis, Zaslavsky et al. (2019) show that color naming is shaped by two major forces: perceptual structure and communicative needs (see also Gibson et al., 2017; Regier, Kay, & Khetarpal, 2007).

With respect to the first question, studies like Winawer et al. (2007) suggest that color terms indeed affect color perception (but see Brogaard & Gatzia, 2017; Firestone & Scholl, 2016; Raftopoulos, 2015, 2017, for general, theory-

driven counter-arguments against any top-down influence of cognition on perception). This claim from Winawer et al. (2007) is supported by several other studies testing cross-linguistic differences (e.g., languages with “two blues” against languages with “one blue”, Greek vs. English: Thierry, Athanasopoulos, Wiggett, Dering, & Kuipers, 2009; Spanish spoken in Uruguay vs. Spanish spoken in Spain: González-Perilli, Rebollo, Maiche, & Arévalo, 2017). In addition, neuroscientific studies provide electrophysiological and neuro-imaging evidence for the claim that language affects perception (e.g., Tan et al., 2008; Thierry et al., 2009, see also Maier & Rahman, 2019). With respect to neurological mechanisms, it has been claimed that Whorfian effects are stronger in the right visual field than in the left visual field (Gilbert, Regier, Kay, & Ivry, 2006, 2008; Regier & Kay, 2009). Since the visual fields project contralaterally to the brain, the right visual field projects to the left hemisphere, which is known for being dominant in language processing. Hence, stronger Whorfian effects in the left hemisphere support the notion that language affects perception.

However, using carefully designed experiments, Witzel and Gegenfurtner (2011) failed to replicate the claimed lateralization. Furthermore Wright, Davies, and Franklin (2015) failed to replicate cross-linguistic Whorfian effects on color memory. To reconcile this mixed evidence, Cibelli, Xu, Austerweil, Griffiths, and Regier (2016) proposed a Bayesian model for Whorfian effects (see also Regier & Xu, 2017). This probabilistic model implements a dual-code representation for color perception: a fine-grained perceptual representation and a coarse-grained linguistic representation. The model assumes that these two representations interact with each other, successfully accommodating observed Whorfian effects. Crucially, the model allows to weight the influence of language, addressing the mixed evidence: The more certain the perceptual information, the lower the effect of language (and vice versa).


The model by Cibelli et al. (2016) is based on the influential *Category Adjustment* model by Huttenlocher, Hedges, and Vevea (2000). Highly similar models were proposed for color perception (Bae, Olkkonen, Allred, & Flombaum, 2015; Witzel, Olkkonen, & Gegenfurtner, 2018) and vowel perception (Feldman, Griffiths, & Morgan, 2009). In addition, these probabilistic models are compatible with the *label-feedback hypothesis* by Lupyan (2012), which in turn is in line with the more general *predictive coding* approach in cognitive science (Lupyan & Clark, 2015). Based on the above reviewed studies, I argue that language affects color perception and that this effect likely operates via probabilistic inference.

REFERENCES (THESIS 3: COLOR PERCEPTION)

- Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*(4), 744–763. <https://doi.org/10.1037/xge0000076>
- Boroditsky, L. (2006). Linguistic relativity. In L. Nadel (Ed.), *Encyclopedia of cognitive science*. <https://doi.org/10.1002/0470018860.s00567>
- Brogaard, B., & Gatzia, D. E. (2017). Is color experience cognitively penetrable? *Topics in Cognitive Science*, *9*, 193–214. <https://doi.org/10.1111/tops.12221>
- Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., & Regier, T. (2016). The Sapir-Whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PLoS ONE*, *11*(7), e0158725. <https://doi.org/10.1371/journal.pone.0158725>
- Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782. <https://doi.org/10.1037/a0017196>
- Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the

- evidence for “top-down” effects. *Behavioral and Brain Sciences*, 39, e229. <https://doi.org/10.1017/S0140525X15000965>
- Gibson, E., Futrell, R., Jara Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40), 10785–10790. <https://doi.org/10.1073/pnas.1619666114>
- Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences*, 103(2), 489–494. <https://doi.org/10.1073/pnas.0509868103>
- Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2008). Support for lateralization of the Whorf effect beyond the realm of color discrimination. *Brain and Language*, 105(2), 91–98. <https://doi.org/10.1016/j.bandl.2007.06.001>
- González-Perilli, F., Rebollo, I., Maiche, A., & Arévalo, A. (2017). Blues in two different spanish-speaking populations. *Frontiers in Communication*, 2, 18. <https://doi.org/10.3389/fcomm.2017.00018>
- Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, 129(2), 220–241. <https://doi.org/10.1037//0096-3445.129.2.220>
- Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist*, 86(1), 65–79. <https://doi.org/10.1525/aa.1984.86.1.02a00050>
- Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, 3, 54. <https://doi.org/10.3389/fpsyg.2012.00054>
- Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science*, 24(4), 279–284. <https://doi.org/10.1177/0963721415570732>
- Maier, M., & Rahman, R. A. (2019). No matter how: Top-down effects of verbal and semantic category knowledge on early visual perception. *Cognitive, Affective, & Behavioral Neuroscience*. <https://doi.org/10.3758/s13415-018-00679-8>
- Raftopoulos, A. (2015). The cognitive impenetrability of perception and theory-ladenness. *Journal for General Philosophy of Science*, 46(1), 87–103. <https://doi.org/10.1007/s10838-015-9288-6>
- Raftopoulos, A. (2017). Pre-cueing, the epistemic role of early vision, and the cognitive impenetrability of early vision. *Frontiers in Psychology*, 8, 1156. <https://doi.org/10.3389/fpsyg.2017.01156>
- Regier, T., & Kay, P. (2009). Language, thought, and color: Whorf was half right. *Trends in Cognitive Sciences*, 13(10), 439–446. <https://doi.org/10.1016/j.tics.2009.07.001>
- Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104(4), 1436–1441. <https://doi.org/10.1073/pnas.0610341104>
- Regier, T., & Xu, Y. (2017). The Sapir-Whorf hypothesis and inference under uncertainty. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(6), e1440. <https://doi.org/10.1002/wcs.1440>
- Tan, L. H., Chan, A. H., Kay, P., Khong, P.-L., Yip, L. K., & Luke, K.-K. (2008). Language affects patterns of brain activation associated with perceptual decision. *Proceedings of the National Academy of Sciences*, 105(10), 4004–4009. <https://doi.org/10.1073/pnas.0800055105>
- Thierry, G., Athanasopoulos, P., Wiggert, A., Dering, B., & Kuipers, J.-R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National Academy of Sciences*, 106(11), 4567–4570. <https://doi.org/10.1073/pnas.0811155106>
- Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19), 7780–7785. <https://doi.org/10.1073/pnas.0701644104>
- Witzel, C. (2018). Misconceptions about colour categories. *Review of Philosophy and Psychology*. <https://doi.org/10.1007/s13164-018-0404-5>
- Witzel, C., & Gegenfurtner, K. R. (2011). Is there a lateralized category effect for color? *Journal of Vision*, 11(12), 16. <https://doi.org/10.1167/11.12.16>
- Witzel, C., Olkkonen, M., & Gegenfurtner, K. R. (2018). A Bayesian model of the memory colour effect. *i-Perception*, 9(3), 1–16. <https://doi.org/10.1177/2041669518771715>
- Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3), 253–265. <https://doi.org/10.1002/wcs.104>
- Wright, O., Davies, I. R., & Franklin, A. (2015). Whorfian effects on colour memory are not reliable. *The Quarterly Journal of Experimental Psychology*, 68(4), 745–758. <https://doi.org/10.1080/17470218.2014.966123>
- Zaslavsky, N., Kemp, C., Tishby, N., & Regier, T. (2019). Color naming reflects both perceptual structure and communicative need. *Topics in Cognitive Science*, 11, 207–219. <https://doi.org/10.1111/tops.12395>

IMAGE CREDITS

- Figures 1.1b, 4.11, 4.12, 4.13, 4.14, 4.15, 4.16, 4.17, 4.20, 5.1, 5.2, 5.4, 5.5, 5.6, 5.7, and 5.8 appeared first in Kluth, Burigo, Schultheis, and Knoeferle (2019), Does direction matter? Linguistic asymmetries reflected in visual attention, *Cognition* 185, 91-120, <https://doi.org/10.1016/j.cognition.2018.09.006>.
- Figure 3.2c: Reprinted from *Cognition* 55 (1), William G. Hayward and Michael Tarr, Spatial language and spatial representation, page 50, Copyright 1995, with permission from Elsevier.
- Figures 5.9 & 5.10 and Table 5.1 appeared first in Kluth and Schultheis (2018), licensed under a Creative Commons Attribution 4.0 International License  (<https://creativecommons.org/licenses/by/4.0/>). They have been slightly modified to follow the format requirements of this thesis.
- Figures 6.1a and 6.1b: Reprinted from *Cognition* 93 (2), Janellen Huttenlocher, Larry V. Hedges, Bryce Corrigan, and L. Elizabeth Crawford, Spatial categories and the estimation of location, pages 77 & 84, Copyright 2004, with permission from Elsevier.

LIST OF FIGURES

Figure 1.1	Schematic representations of (a) the height component and (b) the angular component of the AVS model.	14
Figure 3.1	Vector end points in (a) the AVS model and (b) the rAVS models (rAVS _{prox} : loosely dashed, rAVS _{comb} : dotted, rAVS _{w-comb} : densely dashed, rAVS _{c-o-m} : solid). The points F ₁ and F ₂ in (c) are AVS' attentional focus points for LOs L ₁ and L ₂ . These are used in the rAVS _{prox} model (solid vectors) as well as in the rAVS _{comb} and rAVS _{w-comb} models.	35
Figure 3.2	Layout of experimental displays and displays used for model simulations for (a, b) Logan and Sadler (1996, exp. 2, <i>above</i>) and for (c, d) Hayward and Tarr (1995, exp. 2, <i>above</i>). For (b, d): LOs are displayed as circles for visualization purposes. The simulations used single point LOs (at the center of the circles). Backgrounds depict rAVS _{w-comb} 's spatial template (lighter color coding higher rating) computed with best fitting parameters for the corresponding data set. For (d): No measurements were reported in Hayward and Tarr (1995), so the same distances as for the Logan and Sadler (1996) data were used. Only LO positions above the RO are considered, because <i>above</i> ratings for positions below the LO were not reported.	46
Figure 3.3	GOF and SHO results for data from (a) Logan and Sadler (1996, exp. 2, <i>above</i>) and (b) Hayward and Tarr (1995, exp. 2, <i>above</i>). Note the different y-axes. Error bars depict 95% confidence intervals of SHO median or mean respectively.	47
Figure 3.4	Displays used for simulating the stimuli from (a, b) exp. 1 and (c, d) exp. 2 from Regier and Carlson (2001, reconstructed from data provided by Regier & Carlson, 2001). LOs are displayed as circles for visualization purposes. The simulations used single point LOs (at the center of the circles). For critical manipulation see Figure 3.5b. ROs are patterned for visualization purposes only. Backgrounds depict rAVS _{w-comb} 's spatial templates (lighter color coding higher rating) computed with best fitting parameters for data from Regier and Carlson (2001, exps. 1–3).	49

- Figure 3.5 Examples of one LO placement for (a) exp. 1 and (b) exp. 2 from Regier and Carlson (2001). The two rectangular ROs that were used in the experiments are overlaid (filled with different patterns) to contrast the effect of the LO placement on the proximal (dashed lines) vs. center-of-mass orientation (solid lines). The fill pattern of the LO depicts with which RO the LO was shown. 50
- Figure 3.6 GOF and SHO results for fitting data from Regier and Carlson (2001, exp. 1): (a) tall rectangle, (b) wide rectangle. Error bars depict 95% confidence intervals of SHO median or mean respectively. 53
- Figure 3.7 GOF and SHO results for fitting data from Regier and Carlson (2001, exp. 2): (a) tall rectangle, (b) wide rectangle. Error bars depict 95% confidence intervals of SHO median or mean respectively. 53
- Figure 3.8 GOF and SHO results for fitting data from Regier and Carlson (2001, exps. 1 & 2, both ROs). Error bars depict 95% confidence intervals of SHO median or mean respectively. 54
- Figure 3.9 Displays used for simulating the stimuli of exp. 3 from Regier and Carlson (2001, reconstructed from data provided by Regier & Carlson, 2001). LOs are displayed as circles for visualization purposes. The simulations used single point LOs (at the center of the circles). Critical manipulation: In the regions directly above the ROs (marked with the dashed boxes), the proximal orientation is constant while the center-of-mass orientation varies. ROs are patterned for visualization purposes only. Backgrounds depict $rAVS_{w-comb}$'s spatial templates (lighter color coding higher rating) computed with best fitting parameters for data from Regier and Carlson (2001, exps. 1–3). 55
- Figure 3.10 GOF and SHO results for fitting (Regier & Carlson, 2001, exp. 3). Error bars depict 95% confidence intervals of SHO median or mean respectively. 57
- Figure 3.11 GOF and SHO results for fitting critical data points from Regier and Carlson (2001, exp. 3). Error bars depict 95% confidence intervals of SHO median or mean respectively. 57
- Figure 3.12 Qualitative comparison for Regier and Carlson (2001, exp. 3, wide rectangle, upper row of LOs) computed with (a) parameter values used by Regier and Carlson (2001) and (b) parameter values of my best fit to the critical data from the wide RO of exp. 3. 58

- Figure 3.13 Displays used for simulating the stimuli of exp. 4 from Regier and Carlson (2001, reconstructed from data provided by Regier & Carlson, 2001). LOs are displayed as circles for visualization purposes. The simulations used single point LOs (at the center of the circles). Critical manipulation: "Point A is above the center of mass of the triangle, Point B is above the midpoint of the base of the triangle, and Point C is placed so that its distance from B equals the distance between A and B" (Regier & Carlson, 2001, p. 285). ROs are patterned for visualization purposes only. Backgrounds depict $rAVS_{w-comb}$'s spatial templates (lighter color coding higher rating) computed with best fitting parameters for data from Regier and Carlson (2001, exp. 4). 59
- Figure 3.14 GOF and SHO results for fitting Regier and Carlson (2001, exp. 4, both ROs). Error bars depict 95% confidence intervals of SHO median or mean respectively. 60
- Figure 3.15 Qualitative comparison for data from Regier and Carlson (2001, exp. 4): (a, b) upright triangle, (c, d) inverted triangle. Computed with (a, c) parameter values used by Regier and Carlson (2001) and (b, d) parameter values of my best fit to the corresponding data. Error bars for the empirical data depict the ± 0.3 difference needed for significance (based on 95% confidence intervals) as reported by Regier and Carlson (2001, p. 285). 61
- Figure 3.16 Displays used for simulating the stimuli of exps. 5 & 6 from Regier and Carlson (2001, reconstructed from data provided by Regier & Carlson, 2001). LOs are displayed as circles for visualization purposes. The simulations used single point LOs (at the center of the circles). Critical manipulation: LOs placed above or below the grazing line (horizontal white dashed line). ROs are patterned for visualization purposes only. Backgrounds depict $rAVS_{w-comb}$'s spatial templates (lighter color coding higher rating) computed with best fitting parameters for data from Regier and Carlson (2001, exps. 5 & 6). 63
- Figure 3.17 GOF and SHO results for fitting Regier and Carlson (2001, exp. 5). Error bars depict 95% confidence intervals of SHO median or mean respectively. 65
- Figure 3.18 GOF and SHO results for fitting Regier and Carlson (2001, exp. 6). Error bars depict 95% confidence intervals of SHO median or mean respectively. 65

- Figure 3.19 Qualitative comparison for Regier and Carlson (2001, exp. 6, critical LOs) computed with (a) parameter values used by Regier and Carlson (2001) and (b) parameter values of my best fit to all data from exp. 6. 66
- Figure 3.20 Displays used for simulating the stimuli of exp. 7 from Regier and Carlson (2001, reconstructed from data provided by Regier & Carlson, 2001). LOs are displayed as circles for visualization purposes. The simulations used single point LOs (at the center of the circles). Critical manipulation: distance of LOs to RO. The dashed box frames the critical points. RO is patterned for visualization purposes only. Background depicts $rAVS_{w-comb}$'s spatial templates (lighter color coding higher rating) computed with best fitting parameters for data from Regier and Carlson (2001, exp. 7). 67
- Figure 3.21 Qualitative comparison for Regier and Carlson (2001, exp. 7): (a, b) upper row of LOs, (c, d) lower row of LOs. Generated with (a, c) parameter values used by Regier and Carlson (2001) and (b, d) parameter values of my best fit to all data from exp. 7. 69
- Figure 3.22 GOF and SHO results for fitting Regier and Carlson (2001, exp. 7). Error bars depict 95% confidence intervals of SHO median or mean respectively. 71
- Figure 3.23 GOF and SHO results for fitting Regier and Carlson (2001, exps. 3 & 7, same RO). Error bars depict 95% confidence intervals of SHO median or mean respectively. 71
- Figure 3.24 GOF and SHO results for fitting the data from all experiments from Regier and Carlson (2001). Error bars depict 95% confidence intervals of SHO median or mean respectively. 72
- Figure 4.1 Spatial configurations to test the effect of relative distance of the LO to the RO. ROs are filled with different patterns for visualization purposes only. 77
- Figure 4.2 Qualitative comparison of (a) $rAVS_{w-comb}$ -generated ratings and (b) AVS-generated ratings for LOs above thin and thick rectangle (see Figure 4.1c). For data generation, I have used model parameters from best fit to all data from Regier and Carlson (2001, Table 3.8). 78
- Figure 4.3 Spatial configurations to test the effect of asymmetrical ROs. (a, b): "L" RO, (c, d): "C" RO. (a, c): two critical LO positions, (b, d): all simulated LO positions with critical LO-pairs connected with a dashed line. \times : center-of-mass of RO. 79

- Figure 4.4 Qualitative comparison of (a) $rAVS_{w-comb}$ -generated ratings and (b) AVS-generated ratings for critical LOs above L and C RO (see Figures 4.3b and 4.3d). For data generation, I have used model parameters from best fit to all data from Regier and Carlson (2001, Table 3.8). 80
- Figure 4.5 ROs and LOs used as input for the PSP method. First comparison was between the two LOs above the C RO, second comparison between the two LOs above the L RO, third comparison between the two LOs above the thin vs. the tall rectangle. 83
- Figure 4.6 PSP results: Estimations of relative volumes in parameter spaces of the models covered by distinct qualitative patterns for spatial configurations depicted in Figure 4.5. Subfigure (d) shows legend for all plots (a)–(c). First symbol in pattern: rating difference for LOs above C RO; second symbol: rating difference for LOs above L RO; third symbol: rating difference for LOs above thin vs. tall rectangle. Two mean ratings were considered to be different if they differed by more than (a) $t_e = 0.1$ (b) $t_e = 0.5$ or (c) $t_e = 1.0$. Mean estimates of three PSP runs are plotted. 84
- Figure 4.7 All ROs and their code names used in the empirical rating study. 86
- Figure 4.8 LO placements with row and column coding for two example ROs. Rows R1–R5 were presented with *über* (*above*), rows R6–R10 were presented with *unter* (*below*). 87
- Figure 4.9 Schematic visualization of a single experimental trial. Not to scale. 88
- Figure 4.10 Empirical rating distributions and fits of Bayesian ordinal regression models (computed with 100 samples from the posterior distribution) visualizing the effect of (a) the preposition and (b) the grazing line. Both Bayesian models were instantiated with prior information from earlier research. 92

- Figure 4.11 Individual *über* (*above*) and *unter* (*below*) acceptability ratings for LOs (not depicted) around the asymmetrical C and mC ROs. Individual acceptability ratings are color-coded (the darker the color, the higher the rating) and plotted near the location of the corresponding LO (to avoid overplotting). LOs (not shown in the visualization) were black circles with a 10-pixel diameter and placed at the intersection of the grid lines. LOs in rows R1–R5 were presented with *über* (*above*), LOs in rows R6–R10 were presented with *unter* (*below*). Only one RO and one LO was visible at a time. For each RO: Dashed line is the bounding box, × is the center-of-mass, ○ is the center-of-object. Neither of the centers nor the bounding box were visible to the participants. Image copyright: See Appendix E. 94
- Figure 4.12 Individual *über* (*above*) and *unter* (*below*) acceptability ratings for LOs (not depicted) around the asymmetrical L and mL ROs. Individual acceptability ratings are color-coded (the darker the color, the higher the rating) and plotted near the location of the corresponding LO (to avoid overplotting). LOs (not shown in the visualization) were black circles with a 10-pixel diameter and placed at the intersection of the grid lines. LOs in rows R1–R5 were presented with *über* (*above*), LOs in rows R6–R10 were presented with *unter* (*below*). Only one RO and one LO was visible at a time. For each RO: Dashed line is the bounding box, × is the center-of-mass, ○ is the center-of-object. Neither of the centers nor the bounding box were visible to the participants. Image copyright: See Appendix E. 95
- Figure 4.13 Empirical rating distributions and fit of Bayesian ordinal regression model (computed with 100 samples from the posterior distribution) contrasting ratings for the two subsets “cavity” (columns C4 & C5, all asymmetrical ROs) and “mass” (columns C2 & C3 for ROs C and L; columns C6 & C7 for ROs mC and mL). Bayesian regression model was computed with brms’s default prior. Image copyright: See Appendix E. 97

- Figure 4.14 Individual *über* (*above*) acceptability ratings for LOs (not depicted) above the thin, the thick, the square, and the tall rectangle. Individual acceptability ratings are color-coded (the darker the color, the higher the rating) and plotted near the location of the corresponding LO (to avoid overplotting). LOs (not shown in the visualization) were black circles with a 10-pixel diameter and placed at the intersection of the grid lines. Only one RO and one LO was visible at a time. Image copyright: See Appendix E. 99
- Figure 4.15 Individual *unter* (*below*) acceptability ratings for LOs (not depicted) below the thin, the thick, the square, and the tall rectangle. Individual acceptability ratings are color-coded (the darker the color, the higher the rating) and plotted near the location of the corresponding LO (to avoid overplotting). LOs (not shown in the visualization) were black circles with a 10-pixel diameter and placed at the intersection of the grid lines. Only one RO and one LO was visible at a time. Image copyright: See Appendix E. 100
- Figure 4.16 Empirical rating distributions and fit of Bayesian ordinal regression model (computed with 100 samples from the posterior distribution) contrasting ratings for the four rectangular ROs. Bayesian regression model was computed with prior distributions supporting higher ratings for taller rectangles. Image copyright: See Appendix E. 101
- Figure 4.17 Visualization of effects of (a) proximal orientation and (b) center-of-mass orientation on acceptability rating as estimated by a Bayesian regression model with these two predictors plus relative distance. Plots treat outcome variable as metric (for visualization purposes) which is not how the ordinal regression model deals with the data. Predictors not on the x-axis were kept constant on meaningful values: Relative distance is constant on its mean values for LOs around each of the four ROs, center-of-mass orientation is constant on mean values for LOs around the thin or tall rectangle, proximal orientation is constant on non-deviating orientation (columns C2–C7) and the mean value of deviating proximal orientation (columns C1 & C8). Little black bars on the x-axis denote actually tested data points. Shaded areas denote 95% CIs. Image copyright: See Appendix E. 103

- Figure 4.18 Heatmap visualizations depicting the number of fixations inside BBs of ROs, separated by RO and preposition. Coordinates are in pixels, starting to count from top left of each RO. Computed with 50×50 bins. 106
- Figure 4.19 Heatmap visualizations depicting the number of fixations inside BBs of ROs, separated by RO and preposition. Coordinates are normalized by the dimensions of each BB such that they are relative to each BB. Computed with 50×50 bins. 107
- Figure 4.20 Heatmap visualizations depicting the number of fixations inside BBs of ROs, separated by the column of the LO and preposition. Coordinates are normalized by the dimensions of each BB such that they are relative to each BB. Computed with 50×50 bins. Image copyright: See Appendix E. 108
- Figure 5.1 Goodness-of-fit (GOF) and simple hold-out (SHO) results for (a)–(c) the data from the study presented in Chapter 4 (collapsing across *über*, *above*, and *unter*, *below*) and (d) data from Regier and Carlson (2001). Error bars show bootstrapped 95% confidence intervals of the SHO medians. Image copyright: See Appendix E. 119
- Figure 5.2 Results of the second PSP analysis: Estimations of relative volumes in parameter spaces of the models covered by distinct qualitative patterns (averaged over three PSP runs). First digit codes for difference in mean *über* (*above*) ratings for 28 LOs above the thin rectangle vs. the tall rectangle. Second digit codes for difference in mean *über* (*above*) ratings for 6 LOs to the left vs. to the right of the center-of-mass of the L-shaped RO. Mean ratings were considered equal if they differed less than (a) $t_e = 0.1$ or (b) $t_e = 0.5$. Image copyright: See Appendix E. 123

- Figure 5.3 Results of the Model Flexibility Analysis (MFA). The lower ϕ , the less flexible is the model. Note the different y-axes. Panels (a) and (c) show ϕ_1 , i.e., results computed with the number of data space cells as suggested by Veksler et al. (2015). Panels (b) and (d) show ϕ_{n2} , i.e., results computed with as many cells for every data-space dimension as there were rating intervals (i.e., 9 for the stimuli from Chapter 4, 10 for stimuli from Regier & Carlson, 2001, which are abbreviated as R&C stimuli in the plots) and normalized by dividing with the corresponding ϕ_{2max} . See Tables C.1 and C.2 for more results. For panels (c) and (d), I used smaller parameter ranges (see Equations 5.4–5.6) to address parts the MFA-critique by Evans et al. (2017). 129
- Figure 5.4 Landscaping results contrasting the rAVS-CoO model with the AVS-BB model on the asymmetrical ROs (collapsing across *über, above, and unter, below*). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 5.1c). Image copyright: See Appendix E. 134
- Figure 5.5 Landscaping results contrasting the rAVS-CoO model with the AVS-BB model on the stimuli from Regier and Carlson (2001). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 5.1d). Image copyright: See Appendix E. 134
- Figure 5.6 Landscaping results contrasting the rAVS_{w-comb} model with the rAVS-CoO model on the whole stimuli set (collapsing across *über, above, and unter, below*). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 5.1a). Image copyright: See Appendix E. 135
- Figure 5.7 Landscaping results contrasting the rAVS_{w-comb} model with the rAVS-CoO model on the asymmetrical ROs only (collapsing across *über, above, and unter, below*). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 5.1c). Image copyright: See Appendix E. 135
- Figure 5.8 Landscaping results contrasting the rAVS_{w-comb} model with the AVS model on the stimuli from Regier and Carlson (2001). The asterisks in (b) and (c) depict the fit to the empirical data (cf. GOFs in Figure 5.1d). Image copyright: See Appendix E. 137

- Figure 5.9 Example experimental display to illustrate the model extension that enables the simulation of rating distributions, fits of extended models, and empirical rating distributions. Image copyright: See Appendix E. 141
- Figure 5.10 Marginal posterior distributions for the rAVS-CoO+ model given rating data from asymmetrical ROs and “uninformative” prior distributions (uniform distributions). Image copyright: See Appendix E. 146
- Figure 6.1 Visualization of stimuli and findings from location estimation studies by Huttenlocher et al. (2004, 1991); Lipinski, Simmering, et al. (2010). Image copyright: See Appendix E. 159

LIST OF TABLES

Table 3.1	Parameter values, correlation coefficients, and nRMSE of best fits to data from Logan and Sadler (1996, exp. 2, <i>above</i>). λ values for rAVS models are presented in parentheses because they do not change the model outcome (see page 34). 48
Table 3.2	Linear model fits relating the empirical data from exps. 1 and 2 from Regier and Carlson (2001) with model-generated data for the same stimuli. I computed model-generated data with parameter values from the best fit to Logan and Sadler (1996, exp. 2, <i>above</i>) shown in Table 3.1 – except for lines denoted with “RC-LS fit” where I used parameter values from the AVS model fit to Logan and Sadler (1996, exp. 2, <i>above</i>) as reported in Regier and Carlson (2001). 51
Table 3.2	Continued: Linear models for exps. 1 and 2 from Regier and Carlson (2001). 52
Table 3.3	Linear model fits relating the empirical data from exp. 3 from Regier and Carlson (2001) with model-generated data for the same stimuli. I computed model-generated data with parameter values from the best fit to Logan and Sadler (1996, exp. 2, <i>above</i>) shown in Table 3.1 – except for lines denoted with “RC-LS fit” where I used parameter values from the AVS model fit to Logan and Sadler (1996, exp. 2, <i>above</i>) as reported in Regier and Carlson (2001). 56
Table 3.4	Linear model fits relating the empirical data from exp. 4 from Regier and Carlson (2001) with model-generated data for the same stimuli. I computed model-generated data with parameter values from the best fit to Logan and Sadler (1996, exp. 2, <i>above</i>) shown in Table 3.1 – except for lines denoted with “RC-LS fit” where I used parameter values from the AVS model fit to Logan and Sadler (1996, exp. 2, <i>above</i>) as reported in Regier and Carlson (2001). 60
Table 3.5	The effect of the grazing line for exps. 5 and 6 using parameters for fitting Logan and Sadler (1996, exp. 2, <i>above</i> , Table 3.1). 64

Table 3.6	Linear model fits relating the empirical data from exps. 5 and 6 from Regier and Carlson (2001) with model-generated data for the same stimuli. I computed model-generated data with parameter values from the best fit to Logan and Sadler (1996, exp. 2, <i>above</i>) shown in Table 3.1 – except for lines denoted with “RC-LS fit” where I used parameter values from the AVS model fit to Logan and Sadler (1996, exp. 2, <i>above</i>) as reported in Regier and Carlson (2001). 64
Table 3.7	Linear model fits relating the empirical data from exp. 7 from Regier and Carlson (2001) with model-generated data for the same stimuli. I computed model-generated data with parameter values from the best fit to Logan and Sadler (1996, exp. 2, <i>above</i>) shown in Table 3.1 – except for lines denoted with “RC-LS fit” where I used parameter values from the AVS model fit to Logan and Sadler (1996, exp. 2, <i>above</i>) as reported in Regier and Carlson (2001). 68
Table 3.8	Model parameter values, nRMSE and correlation of the best fit to all data from Regier and Carlson (2001). λ values for rAVS models are presented in parentheses because they do not change the model outcome (see page 34). 73
Table 4.1	Absolute and relative number of fixations (a) inside the bounding boxes of the ROs (leftmost column), (b) split by left or right landing positions (left part of table), and (b) close to the center-of-mass or center-of-object of the RO (no more than 25 pixel in x or y direction, right part of table). * = For rectangular ROs, center-of-mass and center-of-object coincide. For these ROs, the numbers are the total number of fixations to their center. 109
Table 5.1	Example input for the cross-match test (Rosenbaum, 2005). Each row describes the response of one subject (empirical or model-generated), each column describes the response to a stimulus (e.g., the left or right LO from Fig. 5.9a). Table copyright: See Appendix E. 144

- Table C.1 Results of the Model Flexibility Analysis (MFA) computed using 50^4 model predictions and parameter ranges in Equations 3.15–3.17 (page 42) and 5.1–5.2 (page 118). The lower ϕ , the less flexible is the model. ϕ_1 follows suggestion by Veksler et al. (2015) to split each dimension of the data space into $\sqrt[3]{50^4}$ cells. ϕ_2 uses the range of the rating scale as domain specific number of cells per data space dimension. ϕ_{n2} normalizes ϕ_2 by dividing with maximal possible ϕ_{2max} . ϕ_1 and ϕ_{n2} are plotted in Figures 5.3a and 5.3b (page 129). 177
- Table C.2 Results of the Model Flexibility Analysis (MFA) computed using 50^4 model predictions and smaller parameter ranges in Equations 5.4–5.6 (page 127). The lower ϕ , the less flexible is the model. ϕ_1 follows suggestion by Veksler et al. (2015) to split each dimension of the data space into $\sqrt[3]{50^4}$ cells. ϕ_2 uses the range of the rating scale as domain specific number of cells per data space dimension. ϕ_{n2} normalizes ϕ_2 by dividing with maximal possible ϕ_{2max} . ϕ_1 and ϕ_{n2} are plotted in Figures 5.3c and 5.3d (page 129). 178

BIBLIOGRAPHY

- Altmann, G. (2007). Journal policies and procedures. *Cognition*, 102(1), 1–6. <https://doi.org/10.1016/j.cognition.2006.11.001>
- Amorapanth, P., Kranjec, A., Bromberger, B., Lehet, M., Widick, P., Woods, A. J., ... Chatterjee, A. (2012). Language, perception, and the schematic representation of spatial relations. *Brain and Language*, 120(3), 226–236. <https://doi.org/10.1016/j.bandl.2011.09.007>
- Anderson, C. H., Van Essen, D. C., & Olshausen, B. A. (2005). Directed visual attention and the dynamic control of information flow. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 11–17). Burlington: Academic Press. <https://doi.org/10.1016/B978-012375731-9/50007-0>
- Arbib, M. A. (2017). Dorsal and ventral streams in the evolution of the language-ready brain: Linking language to the world. *Journal of Neurolinguistics*, 43, Part B, 228–253. <https://doi.org/10.1016/j.jneuroling.2016.12.003>
- Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79. <https://doi.org/10.1214/09-SS054>
- Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(4), 723–742.
- Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., ... Kumaran, D. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557, 429–433. <https://doi.org/10.1038/s41586-018-0102-6>
- Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, 59(1), 617–645. <https://doi.org/10.1146/annurev.psych.59.103006.093639>
- Bloom, P., Peterson, M. A., Nadel, L., & Garret, M. F. (Eds.). (1996). *Language and space*. Cambridge, Massachusetts; London, England: MIT Press.
- Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75(1), 1–28. [https://doi.org/10.1016/S0010-0277\(99\)00073-6](https://doi.org/10.1016/S0010-0277(99)00073-6)
- Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799.
- Brouwer, A.-M., Franz, V. H., & Gegenfurtner, K. R. (2009). Differences in fixations between grasping and viewing objects. *Journal of Vision*, 9(1), 18–18. <https://doi.org/10.1167/9.1.18>
- Brown, P., & Levinson, S. C. (1993). *Linguistic and nonlinguistic coding of spatial arrays: Explorations in Mayan cognition*. Working paper 24. Cognitive Anthropology Research Group, Max Planck Institute for Psycholinguistics.
- Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, 97(4), 523–547.
- Bundesden, C. (1998). A computational theory of visual attention. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 353(1373), 1271–1281. <https://doi.org/10.1098/rstb.1998.0282>
- Bundesden, C., Vangkilde, S., & Petersen, A. (2015). Recent developments in a computational theory of visual attention (TVA). *Vision Research*, 116, 210–218. <https://doi.org/10.1016/j.visres.2014.11.005>
- Burigo, M., & Coventry, K. (2005). Reference frame conflict in assigning direction to space. In C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, & T. Barkowsky (Eds.), *Spatial cognition IV. Spatial cognition 2004*. (Vol. 3343, pp. 111–123). Berlin: Springer. https://doi.org/10.1007/978-3-540-32255-9_7
- Burigo, M., Coventry, K. R., Cangelosi, A., & Lynott, D. (2016). Spatial language and converseness. *Quarterly Journal of Experimental Psychology*, 69(12), 2319–2337. <https://doi.org/10.1080/17470218.2015.1124894>
- Burigo, M., & Knoeferle, P. (2015). Visual attention during spatial language comprehension. *PLOS ONE*, 10(1), e0115758. <https://doi.org/10.1371/journal.pone.0115758>
- Burigo, M., & Sacchi, S. (2013). Object orientation affects spatial language comprehension. *Cognitive Science*, 37(8), 1471–1492. <https://doi.org/10.1111/cogs.12041>
- Burigo, M., & Schultheis, H. (2018). The effects of direction and orientation of located objects on spatial language comprehension. *Language & Cognition*, 10(2), 298–328. <https://doi.org/10.1017/langcog.2018.3>
- Bush, D., Barry, C., Manson, D., & Burgess, N. (2015). Using grid cells for navigation. *Neuron*, 87(3), 507–520. <https://doi.org/10.1016/j.neuron.2015.07.006>
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Caligiore, D., & Fischer, M. H. (2013). Vision, action and language unified through embodiment. *Psychological Research*, 77(1), 1–6. <https://doi.org/10.1007/s00426-012-0417-0>
- Cangelosi, A. (2010). Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, 7(2), 139–151. <https://doi.org/10.1016/j.plrev.2010.02.001>
- Canty, A., & Ripley, B. (2016). boot: Bootstrap R (S-Plus) Functions [Computer software manual].

- Retrieved from <https://CRAN.R-project.org/package=boot> (R package version 1.3-18)
- Carlson, L. A. (1999). Selecting a reference frame. *Spatial Cognition and Computation*, 1(4), 365–379. <https://doi.org/10.1023/A:1010071109785>
- Carlson, L. A. (2007). Commentary: Linking internal representations to the external world via spatial relations. In J. M. Plumert & J. P. Spencer (Eds.), *The emerging spatial mind* (pp. 248–260). Oxford: Oxford University Press.
- Carlson, L. A., & Logan, G. D. (2001). Using spatial terms to select an object. *Memory & Cognition*, 29(6), 883–892. <https://doi.org/10.3758/BF03196417>
- Carlson, L. A., & Logan, G. D. (2005). Attention and spatial language. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 330–336). Burlington: Academic Press. <https://doi.org/10.1016/B978-012375731-9/50058-6>
- Carlson, L. A., Regier, T., Lopez, W., & Corrigan, B. (2006). Attention unites form and function in spatial language. *Spatial Cognition and Computation*, 6(4), 295–308. https://doi.org/10.1207/s15427633scc0604_1
- Carlson-Radvansky, L. A., Covey, E. S., & Lattanzi, K. M. (1999). “What” effects on “where”: Functional influences on spatial relations. *Psychological Science*, 10(6), 516–521. <https://doi.org/10.1111/1467-9280.00198>
- Carlson-Radvansky, L. A., & Irwin, D. E. (1993). Frames of reference in vision and language: Where is above? *Cognition*, 46(3), 223–244. [https://doi.org/10.1016/0010-0277\(93\)90011-J](https://doi.org/10.1016/0010-0277(93)90011-J)
- Carlson-Radvansky, L. A., & Logan, G. D. (1997). The influence of reference frame selection on spatial template construction. *Journal of Memory and Language*, 37(3), 411–437. <https://doi.org/10.1006/jmla.1997.2519>
- Carlson Radvansky, L. A., & Jiang, Y. (1998). Inhibition accompanies reference-frame selection. *Psychological Science*, 9(5), 386–391. <https://doi.org/10.1111/1467-9280.00072>
- Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, 51(13), 1484–1525. <https://doi.org/10.1016/j.visres.2011.04.012>
- Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, 47(1), 30–49. <https://doi.org/10.1006/jmla.2001.2832>
- Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *WIREs Cognitive Science*, 1(6), 811–823. <https://doi.org/10.1002/wcs.79>
- Coello, Y., & Fischer, M. H. (Eds.). (2015). *Perceptual and emotional embodiment: Foundations of embodied cognition* (Vol. 1). London: Routledge.
- Cohen, E. H., Schnitzer, B. S., Gersch, T. M., Singh, M., & Kowler, E. (2007). The relationship between spatial pooling and attention in saccadic and perceptual tasks. *Vision Research*, 47(14), 1907–1923. <https://doi.org/10.1016/j.visres.2007.03.018>
- Conder, J., Fridriksson, J., Baylis, G. C., Smith, C. M., Boiteau, T. W., & Almor, A. (2017). Bilateral parietal contributions to spatial language. *Brain and Language*, 164, 16–24. <https://doi.org/10.1016/j.bandl.2016.09.007>
- Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3), 201–215. <https://doi.org/10.1038/nrn755>
- Coventry, K. R., Carmichael, R., & Garrod, S. C. (1994). Spatial prepositions, object-specific function, and task requirements. *Journal of Semantics*, 11(4), 289–309. <https://doi.org/10.1093/jos/11.4.289>
- Coventry, K. R., & Garrod, S. C. (2004). *Saying, seeing, and acting: The psychological semantics of spatial prepositions*. Hove and New York: Psychology Press, Taylor and Francis.
- Coventry, K. R., Lynott, D., Cangelosi, A., Monrouxe, L., Joyce, D., & Richardson, D. C. (2010). Spatial language, visual attention, and perceptual simulation. *Brain and Language*, 112(3), 202–213. <https://doi.org/10.1016/j.bandl.2009.06.001>
- Coventry, K. R., Prat Sala, M., & Richards, L. (2001). The interplay between geometry and function in the comprehension of *over*, *under*, *above*, and *below*. *Journal of Memory and Language*, 44(3), 376–398. <https://doi.org/10.1006/jmla.2000.2742>
- Crawford, L. E., Huttenlocher, J., & Hedges, L. V. (2006). Within-category feature correlations and Bayesian adjustment strategies. *Psychonomic Bulletin & Review*, 13(2), 245–250. <https://doi.org/10.3758/BF03193838>
- Crawford, L. E., Regier, T., & Huttenlocher, J. (2000). Linguistic and non-linguistic spatial categorization. *Cognition*, 75(3), 209–235. [https://doi.org/10.1016/S0010-0277\(00\)00064-0](https://doi.org/10.1016/S0010-0277(00)00064-0)
- Cueva, C. J., & Wei, X.-X. (2018). Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In *International Conference on Learning Representations*. Retrieved from <https://openreview.net/forum?id=B17JT0e0>
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29. <https://doi.org/10.1177/0956797613504966>
- Davis, G. J., & Gibson, B. S. (2012). Going rogue in the spatial cuing paradigm: High spatial validity is insufficient to elicit voluntary shifts of attention. *Journal of Experimental Psychology: Human Perception and Performance*, 38(5), 1192–1201. <https://doi.org/10.1037/a0027595>
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press.

- Dawson, M. R. (1988). Fitting the ex-Gaussian equation to reaction time distributions. *Behavior Research Methods, Instruments, & Computers*, 20(1), 54–57. <https://doi.org/10.3758/BF03202603>
- Dent, K. (2009). Coding categorical and coordinate spatial relations in visual–spatial short-term memory. *The Quarterly Journal of Experimental Psychology*, 62(12), 2372–2387. <https://doi.org/10.1080/17470210902853548>
- Desanghere, L., & Marotta, J. J. (2015). The influence of object shape and center of mass on grasp and gaze. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.01537>
- Dessalegn, B., & Landau, B. (2008). More than meets the eye: The role of language in binding and maintaining feature conjunctions. *Psychological Science*, 19(2), 189–195. <https://doi.org/10.1111/j.1467-9280.2008.02066.x>
- Dessalegn, B., & Landau, B. (2013). Interaction between language and vision: It's momentary, abstract, and it develops. *Cognition*, 127, 331–344. <https://doi.org/10.1016/j.cognition.2013.02.003>
- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6(3), 274–290. <https://doi.org/10.1177/1745691611406920>
- Downing, C. J., & Pinker, S. (1985). The spatial structure of visual attention. In M. I. Posner & O. S. M. Marin (Eds.), *Mechanisms of attention: Attention and performance XI* (pp. 171–187). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Dudschig, C., Souman, J., Lachmair, M., de la Vega, I., & Kaup, B. (2013). Reading “sun” and looking up: The influence of language on saccadic eye movements in the vertical dimension. *PLOS ONE*, 8(2), e56872. <https://doi.org/10.1371/journal.pone.0056872>
- Dunn, B., Kamide, Y., & Scheepers, C. (2014). Hearing “moon” and looking up: Word-related spatial associations facilitate saccades to congruent locations. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 433–438). Austin, Texas: Cognitive Science Society.
- Eaton, J. W., Bateman, D., Hauberg, S., & Wehbring, R. (2015). *GNU Octave version 4.0.0 manual: A high-level interactive language for numerical computations*. (<http://www.gnu.org/software/octave/doc/interpreter>)
- Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 123(2), 161–177. <https://doi.org/10.1037/0096-3445.123.2.161>
- Eliasmith, C. (2015). *How to build a brain*. Oxford: Oxford University Press.
- Eliasmith, C., & Anderson, C. H. (2004). *Neural engineering*. Cambridge, Massachusetts; London, England: MIT Press.
- Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, 338(6111), 1202–1205. <https://doi.org/10.1126/science.1225266>
- Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, 25(1), 219–234. <https://doi.org/10.3758/s13423-017-1317-5>
- Evans, N. J., Howard, Z. L., Heathcote, A., & Brown, S. D. (2017). Model flexibility analysis does not measure the persuasiveness of a fit. *Psychological Review*, 124(3), 339–345. <https://doi.org/10.1037/rev0000057>
- Farran, E. K., & O'Leary, B. (2016). Children's ability to bind and maintain colour–location conjunctions: The effect of spatial language cues. *Journal of Cognitive Psychology*, 28(1), 44–51. <https://doi.org/10.1080/20445911.2015.1092980>
- Feist, M. I., & Gentner, D. (2003). Factors involved in the use of in and on. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (pp. 390–395). Boston, Massachusetts: Cognitive Science Society.
- Feist, M. I., & Gentner, D. (2007). Spatial language influences memory for spatial scenes. *Memory & Cognition*, 35(2), 283–296. <https://doi.org/10.3758/BF03193449>
- Fernandez-Duque, D., & Johnson, M. L. (1999). Attention metaphors: How metaphors guide the cognitive psychology of attention. *Cognitive Science*, 23(1), 83–116. https://doi.org/10.1207/s15516709cog2301_4
- Fernandez-Duque, D., & Johnson, M. L. (2002). Cause and effect theories of attention: The role of conceptual metaphors. *Review of General Psychology*, 6(2), 153–165. <https://doi.org/10.1037/1089-2680.6.2.153>
- Fernández-i-Marín, X. (2016). ggmcmc: Analysis of MCMC Samples and Bayesian Inference. *Journal of Statistical Software*, 70(9), 1–20. <https://doi.org/10.18637/jss.v070.i09>
- Fischer, M. H., & Coello, Y. (Eds.). (2015). *Conceptual and interactive embodiment: Foundations of embodied cognition* (Vol. 2). London: Routledge.
- Franciotti, R., D'Ascenzo, S., Di Domenico, A., Onofri, M., Tommasi, L., & Laeng, B. (2013). Focusing narrowly or broadly attention when judging categorical and coordinate spatial relations: A MEG study. *PLOS ONE*, 8(12), e83434. <https://doi.org/10.1371/journal.pone.0083434>
- 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. <https://doi.org/10.1016/j.cognition.2011.11.002>

- Franklin, N., Henkel, L. A., & Zangas, T. (1995). Parsing surrounding space into regions. *Memory & Cognition*, 23(4), 397–407. <https://doi.org/10.3758/BF03197242>
- Fum, D., Del Missier, F., & Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research*, 8(3), 135–142. <https://doi.org/10.1016/j.cogsys.2007.07.001>
- Gapp, K.-P. (1995). An empirically validated model for computing spatial relations. In I. Wachsmuth, C.-R. Rollinger, & W. Brauer (Eds.), *KI-95: Advances in artificial intelligence* (Vol. 981, pp. 245–256). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-60343-3_41
- Garthwaite, P. H., Fan, Y., & Sisson, S. A. (2016). Adaptive optimal scaling of Metropolis-Hastings algorithms using the Robbins-Monro process. *Communications in Statistics – Theory and Methods*, 45(17), 5098–5111. <https://doi.org/10.1080/03610926.2014.936562>
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6), 997–1016. <https://doi.org/10.1007/s11222-013-9416-2>
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. <https://doi.org/10.1214/ss/1177011136>
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155–170. [https://doi.org/10.1016/S0364-0213\(83\)80009-3](https://doi.org/10.1016/S0364-0213(83)80009-3)
- Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195–235). Cambridge, Massachusetts: MIT Press.
- Gentner, D., Özyürek, A., Gürcanlı, Ö., & Goldin-Meadow, S. (2013). Spatial language facilitates spatial cognition: Evidence from children who lack language input. *Cognition*, 127(3), 318–330. <https://doi.org/10.1016/j.cognition.2013.01.003>
- Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, 233(4771), 1416–1419. <https://doi.org/10.1126/science.3749885>
- Gibson, B. S., & Kingstone, A. (2006). Visual attention and the semantics of space: Beyond central and peripheral cues. *Psychological Science*, 17(7), 622–627. <https://doi.org/10.1111/j.1467-9280.2006.01754.x>
- Gibson, B. S., & Sztybel, P. (2014). The spatial semantics of symbolic attention control. *Current Directions in Psychological Science*, 23(4), 271–276. <https://doi.org/10.1177/0963721414536728>
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5), 587–606. <https://doi.org/10.1016/j.socec.2004.09.033>
- Gorniak, P., & Roy, D. (2007). Situated language understanding as filtering perceived affordances. *Cognitive Science*, 31(2), 197–231. <https://doi.org/10.1080/15326900701221199>
- Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 59–100). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511816772.006>
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346. [https://doi.org/10.1016/0167-2789\(90\)90087-6](https://doi.org/10.1016/0167-2789(90)90087-6)
- Hartsuiker, R. J., Huettig, F., & Olivers, C. N. (2011). Visual search and visual world: Interactions among visual attention, language, and working memory (introduction to the special issue). *Acta Psychologica*, 137(2), 135–137. <https://doi.org/10.1016/j.actpsy.2011.01.005>
- Hayward, W. G., & Tarr, M. J. (1995). Spatial language and spatial representation. *Cognition*, 55(1), 39–84. [https://doi.org/10.1016/0010-0277\(94\)00643-Y](https://doi.org/10.1016/0010-0277(94)00643-Y)
- Heller, R., Small, D., & Rosenbaum, P. (2012). crossmatch: The Cross-match Test [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=crossmatch> (R package version 1.3-1)
- Hoekstra, R., Morey, R. D., Rouder, J., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5), 1157–1164. <https://doi.org/10.3758/s13423-013-0572-3>
- Holcombe, A. O., Linares, D., & Vaziri-Pashkam, M. (2011). Perceiving spatial relations via attentional tracking and shifting. *Current Biology*, 21(13), 1135–1139. <https://doi.org/10.1016/j.cub.2011.05.031>
- Hörberg, T. (2008). Influences of form and function on the acceptability of projective prepositions in Swedish. *Spatial Cognition & Computation*, 8(3), 193–218. <https://doi.org/10.1080/13875860801993652>
- Huettig, F., Olivers, C. N., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychologica*, 137(2), 138–150. <https://doi.org/10.1016/j.actpsy.2010.07.013>
- Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, 137(2), 151–171. <https://doi.org/10.1016/j.actpsy.2010.11.003>
- Huttenlocher, J., Hedges, L. V., Corrigan, B., & Crawford, L. E. (2004). Spatial categories and the estimation of location. *Cognition*, 93(2), 75–97. <https://doi.org/10.1016/j.cognition.2003.10.006>
- Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, 98(3), 352–376. <https://doi.org/10.1037/0033-295X.98.3.352>

- Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, *129*(2), 220–241. <https://doi.org/10.1037//0096-3445.129.2.220>
- Huttenlocher, J., & Strauss, S. (1968). Comprehension and a statement's relation to the situation it describes. *Journal of Verbal Learning and Verbal Behavior*, *7*(2), 300–304. [https://doi.org/10.1016/S0022-5371\(68\)80005-2](https://doi.org/10.1016/S0022-5371(68)80005-2)
- Jager, G., & Postma, A. (2003). On the hemispheric specialization for categorical and coordinate spatial relations: A review of the current evidence. *Neuropsychologia*, *41*(4), 504–515. [https://doi.org/10.1016/S0028-3932\(02\)00086-6](https://doi.org/10.1016/S0028-3932(02)00086-6)
- Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169–188. <https://doi.org/10.1017/S0140525X10003134>
- Kemmerer, D. (2006). The semantics of space: Integrating linguistic typology and cognitive neuroscience. *Neuropsychologia*, *44*(9), 1607–1621. <https://doi.org/10.1016/j.neuropsychologia.2006.01.025>
- Kemmerer, D., & Tranel, D. (2000). A double dissociation between linguistic and perceptual representations of spatial relationships. *Cognitive Neuropsychology*, *17*(5), 393–414. <https://doi.org/10.1080/026432900410766>
- Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054. <https://doi.org/10.1126/science.1218811>
- Khetarpal, N., Majid, A., & Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 2396–2401). Austin, Texas: Cognitive Science Society.
- Kiehn, O., & Forssberg, H. (2014). *Scientific background: The brain's navigational place and grid cell system*. http://www.nobelprizemedicine.org/wp-content/uploads/2014/10/Scientific-background_2014.pdf. (retrieved September 7, 2018)
- Kim, W., Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). An MCMC-based method of comparing connectionist models in cognitive science. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference* (pp. 937–944). MIT Press.
- Kluth, T. (2014). *One focus or many? Modeling attentional distributions during spatial term use*. (Master's thesis, Universität Bremen, Fachbereich 3 Mathematik und Informatik). <https://doi.org/10.31237/osf.io/9xu7g>
- Kluth, T. (2018). *A C++ implementation of cognitive models of spatial language understanding as well as pertinent empirical data and analyses*. Bielefeld University. <https://doi.org/10.4119/unibi/2918231>
- Kluth, T., Burigo, M., & Knoeferle, P. (2015). Spatial language comprehension: A computational investigation of the directionality of attention. In A. Gatt & H. Mitterer (Eds.), *Architectures & Mechanisms for Language Processing (AMLaP 2015)* (p. 88). Valetta, Malta.
- Kluth, T., Burigo, M., & Knoeferle, P. (2016a). Investigating the parameter space of cognitive models of spatial language comprehension. In 5. *Interdisziplinärer Workshop Kognitive Systeme: Mensch, Teams, Systeme und Automaten. Verstehen, Beschreiben und Gestalten Kognitiver (Technischer) Systeme*. Bochum, Germany.
- Kluth, T., Burigo, M., & Knoeferle, P. (2016b). Modeling shifts of attention during spatial language comprehension. In T. Tenbrink, A. Foltz, A. Wallington, J. O. Redondo, J. Ryan, & E. Bedford (Eds.), *UK-CLC 2016 Conference Proceedings* (p. 71). Bangor, Wales, UK.
- Kluth, T., Burigo, M., & Knoeferle, P. (2016c). Shifts of attention during spatial language comprehension: A computational investigation. In J. v. d. Herik & J. Filipe (Eds.), *Proceedings of the 8th International Conference on Agents and Artificial Intelligence – Volume 2: ICAART* (pp. 213–222). Rome, Italy: SCITEPRESS. <https://doi.org/10.5220/0005851202130222>
- Kluth, T., Burigo, M., & Knoeferle, P. (2017). Modeling the directionality of attention during spatial language comprehension. In J. v. d. Herik & J. Filipe (Eds.), *Agents and artificial intelligence* (Vol. 10162, pp. 283–301). Cham, Switzerland: Springer International Publishing AG. https://doi.org/10.1007/978-3-319-53354-4_16
- Kluth, T., Burigo, M., Schultheis, H., & Knoeferle, P. (2016a). Distinguishing cognitive models of spatial language understanding. In D. Reitter & F. E. Ritter (Eds.), *Proceedings of the International Conference on Cognitive Modeling* (pp. 230–231). University Park, Pennsylvania: Penn State.
- Kluth, T., Burigo, M., Schultheis, H., & Knoeferle, P. (2016b). The role of the center-of-mass in evaluating spatial language. In T. Barkowsky, H. Schultheis, J. van de Ven, & Z. Falomir Llansola (Eds.), *13th Biannual Conference of the German Society for Cognitive Science: Proceedings* (pp. 11–14). Bremen, Germany.
- Kluth, T., Burigo, M., Schultheis, H., & Knoeferle, P. (2017). Size matters: Effects of relative distance on the acceptability of spatial prepositions. In A. Shestakova et al. (Eds.), *Proceedings of the 10th Embodied and Situated Language Processing Conference* (p. 21). Moscow, Russia: Centre for Cognition and Decision Making, Higher School of Economics.
- Kluth, T., Burigo, M., Schultheis, H., & Knoeferle, P. (2019). Does direction matter? Linguistic asymmetries reflected in visual attention. *Cognition*, *185*, 91–120. <https://doi.org/10.1016/j.cognition.2018.09.006>

- Kluth, T., & Schultheis, H. (2014). Attentional distribution and spatial language. In C. Freksa, B. Nebel, M. Hegarty, & T. Barkowsky (Eds.), *Spatial cognition IX. Spatial cognition 2014*. (Vol. 8684, pp. 76–91). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-11215-2_6
- Kluth, T., & Schultheis, H. (2018). Rating distributions and Bayesian inference: Enhancing cognitive models of spatial language use. In M. Idiart, A. Lenci, T. Poibeau, & A. Villavicencio (Eds.), *Proceedings of the Eighth Workshop on Cognitive Aspects of Computational Language Learning and Processing, co-located with the 56th Annual Meeting of the Association for Computational Linguistics* (pp. 47–55). Melbourne, Australia: Association for Computational Linguistics.
- Knoeferle, P., & Guerra, E. (2016). Visually situated language comprehension. *Language and Linguistics Compass*, 10(2), 66–82. <https://doi.org/10.1111/lnc3.12177>
- Knoeferle, P., Pyykkönen-Klauck, P., & Crocker, M. W. (Eds.). (2016). *Visually situated language comprehension*. Amsterdam, Philadelphia: John Benjamins Publishing Company. <https://doi.org/10.1075/aicr.93>
- Kosslyn, S. M. (1987). Seeing and imagining in the cerebral hemispheres: a computational approach. *Psychological Review*, 94(2), 148–175.
- Kosslyn, S. M. (2006). You can play 20 questions with nature and win: Categorical versus coordinate spatial relations as a case study. *Neuropsychologia*, 44(9), 1519–1523. <https://doi.org/10.1016/j.neuropsychologia.2006.01.022>
- Kosslyn, S. M., Koenig, O., Barrett, A., Cave, C. B., Tang, J., & Gabrieli, J. D. (1989). Evidence for two types of spatial representations: Hemispheric specialization for categorical and coordinate relations. *Journal of Experimental Psychology: Human Perception and Performance*, 15(4), 723–735.
- Kowler, E. (2011). Eye movements: The past 25 years. *Vision Research*, 51(13), 1457–1483. <https://doi.org/10.1016/j.visres.2010.12.014>
- Kranjec, A., Lupyan, G., & Chatterjee, A. (2014). Categorical biases in perceiving spatial relations. *PLOS ONE*, 9(5), e98604. <https://doi.org/10.1371/journal.pone.0098604>
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2), 573–603. <https://doi.org/10.1037/a0029146>
- Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Amsterdam: Academic Press.
- Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, 25(1), 178–206. <https://doi.org/10.3758/s13423-016-1221-4>
- Kullback, S. (1987). Letter to the editor: The Kullback-Leibler distance. *The American Statistician*, 41(4), 340–341. <https://doi.org/10.1080/00031305.1987.10475510>
- LaBerge, D., & Brown, V. (1989). Theory of attentional operations in shape identification. *Psychological Review*, 96(1), 101–124. <https://doi.org/10.1037/0033-295X.96.1.101>
- Laeng, B., Okubo, M., Saneyoshi, A., & Michimata, C. (2011). Processing spatial relations with different apertures of attention. *Cognitive Science*, 35(2), 297–329. <https://doi.org/10.1111/j.1551-6709.2010.01139.x>
- Landau, B. (2017). Update on “What” and “Where” in Spatial Language: A New Division of Labor for Spatial Terms. *Cognitive Science*, 41(S2), 321–350. <https://doi.org/10.1111/cogs.12410>
- Landau, B., & Jackendoff, R. (1993). “What” and “where” in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2), 217–265. <https://doi.org/10.1017/S0140525X00029733>
- Lee, C., Rohrer, W. H., & Sparks, D. L. (1988). Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature*, 332, 357–360. <https://doi.org/10.1038/332357a0>
- Levelt, W. J. (1984). Some perceptual limitations on talking about space. In A. J. v. Doorn & M. A. Bouman (Eds.), *Limits in perception* (pp. 323–358). Utrecht: VNU Science Press.
- Levinson, S. C. (1996). Frames of reference and Molyneux’s question: Crosslinguistic evidence. In P. Bloom, M. Peterson, L. Nadel, & Garrett M. (Eds.), *Language and Space* (pp. 109–169). Cambridge, Massachusetts; London, England: MIT Press.
- Levinson, S. C. (2003). *Space in language and cognition: Explorations in cognitive diversity*. Cambridge: Cambridge University Press.
- Levinson, S. C., Kita, S., Haun, D., & Rasch, B. H. (2002). Returning the tables: Language affects spatial reasoning. *Cognition*, 84(2), 155–188. [https://doi.org/10.1016/S0010-0277\(02\)00045-8](https://doi.org/10.1016/S0010-0277(02)00045-8)
- Li, P., & Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. *Cognition*, 83(3), 265–294. [https://doi.org/10.1016/S0010-0277\(02\)00009-4](https://doi.org/10.1016/S0010-0277(02)00009-4)
- Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79, 328–348. <https://doi.org/10.1016/j.jesp.2018.08.009>
- Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, 15(1), 22–25. <https://doi.org/10.1111/j.1467-9639.1993.tb00252.x>
- Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(6), 1490–1511. <https://doi.org/10.1037/a0022643>
- Lipinski, J., Simmering, V. R., Johnson, J. S., & Spencer, J. P. (2010). The role of experience in loca-

- tion estimation: Target distributions shift location memory biases. *Cognition*, 115(1), 147–153. <https://doi.org/10.1016/j.cognition.2009.12.008>
- Lipinski, J., Spencer, J. P., & Samuelson, L. K. (2009). It's in the eye of the beholder: Spatial language and spatial memory use the same perceptual reference frames. In K. S. Mix, L. B. Smith, & M. Gasser (Eds.), *The spatial foundations of language and cognition* (Vol. 4, pp. 102–131). Oxford: Oxford University Press.
- Lipinski, J., Spencer, J. P., & Samuelson, L. K. (2010a). Biased feedback in spatial recall yields a violation of delta rule learning. *Psychonomic Bulletin & Review*, 17(4), 581–588. <https://doi.org/10.3758/PBR.17.4.581>
- Lipinski, J., Spencer, J. P., & Samuelson, L. K. (2010b). Corresponding delay-dependent biases in spatial language and spatial memory. *Psychological Research*, 74(3), 337–351. <https://doi.org/10.1007/s00426-009-0255-x>
- Livins, K. A., Dumas, L. A., & Spivey, M. J. (2016). Shaping relations: Exploiting relational features for visuospatial priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(1), 127–139. <https://doi.org/10.1037/xlm0000149>
- Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, 50(4), 315–353. <https://doi.org/10.1016/j.cogpsych.2004.09.004>
- Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, 20(5), 1015–1036. <https://doi.org/10.1037/0096-1523.20.5.1015>
- Logan, G. D. (1995). Linguistic and conceptual control of visual spatial attention. *Cognitive Psychology*, 28(2), 103–174. <https://doi.org/10.1006/cogp.1995.1004>
- Logan, G. D. (1996). The CODE theory of visual attention: An integration of space-based and object-based attention. *Psychological Review*, 103(4), 603–649. <https://doi.org/10.1037/0033-295X.103.4.603>
- Logan, G. D., & Bundesen, C. (1996). Spatial effects in the partial report paradigm: A challenge for theories of visual spatial attention. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 35, pp. 243–282). San Diego, California: Academic Press. [https://doi.org/10.1016/S0079-7421\(08\)60577-0](https://doi.org/10.1016/S0079-7421(08)60577-0)
- Logan, G. D., & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.), *Language and space* (pp. 493–530). Cambridge, Massachusetts; London, England: MIT Press.
- Logan, G. D., & Zbrodoff, N. J. (1999). Selection for cognition: Cognitive constraints on visual spatial attention. *Visual Cognition*, 6(1), 55–81. <https://doi.org/10.1080/713756797>
- Lovett, A., & Forbus, K. (2009). Using a visual routine to model the computation of positional relationships. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 1882–1887). Austin, Texas: Cognitive Science Society.
- Madras, N. N. (2002). Markov Chain Monte Carlo. In *Lectures on monte carlo methods* (pp. 53–73). Providence, Rhode Island: American Mathematical Society.
- Majid, A., Bowerman, M., Kita, S., Haun, D., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, 8(3), 108–114. <https://doi.org/10.1016/j.tics.2004.01.003>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman and Company.
- Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, Massachusetts: MIT Press.
- Matin, E., Shao, K., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Attention, Perception, & Psychophysics*, 53(4), 372–380. <https://doi.org/10.3758/BF03206780>
- Mayor, J., Gomez, P., Chang, F., & Lupyan, G. (2014). Connectionism coming of age: Legacy and future challenges. *Frontiers in Psychology*, 5, 187. <https://doi.org/10.3389/fpsyg.2014.00187>
- McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, 1(1), 11–38. <https://doi.org/10.1111/j.1756-8765.2008.01003.x>
- McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Psychological and biological models* (Vol. 2). Cambridge, Massachusetts: MIT Press.
- Melcher, D., & Kowler, E. (1999). Shapes, surfaces and saccades. *Vision Research*, 39(17), 2929–2946. [https://doi.org/10.1016/S0042-6989\(99\)00029-2](https://doi.org/10.1016/S0042-6989(99)00029-2)
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087–1092. <https://doi.org/10.1063/1.1699114>
- Miller, H. E., Patterson, R., & Simmering, V. R. (2016). Language supports young children's use of spatial relations to remember locations. *Cognition*, 150, 170–180. <https://doi.org/10.1016/j.cognition.2016.02.006>
- Miller, H. E., Vlach, H. A., & Simmering, V. R. (2017). Producing spatial words is not enough: Understanding the relation between language and spatial cognition. *Child Development*, 88(6), 1966–

1982. <https://doi.org/10.1111/cdev.12664>
- Moore, K. E. (2014). *The spatial language of time: Metaphor, metonymy, and frames of reference*. Amsterdam, Philadelphia: John Benjamins Publishing Company. <https://doi.org/10.1075/hcp.42>
- Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, 31, 69–89. <https://doi.org/10.1146/annurev.neuro.31.061307.090723>
- Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*, 72, 1–5. <https://doi.org/10.1016/j.jmp.2016.01.002>
- Munnich, E., Landau, B., & Doshier, B. A. (2001). Spatial language and spatial representation: A cross-linguistic comparison. *Cognition*, 81(3), 171–208. [https://doi.org/10.1016/S0010-0277\(01\)00127-5](https://doi.org/10.1016/S0010-0277(01)00127-5)
- Nakayama, K., & Martini, P. (2011). Situating visual search. *Vision Research*, 51(13), 1526–1537. <https://doi.org/10.1016/j.visres.2010.09.003>
- Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, 50(2), 101–122. <https://doi.org/10.1016/j.jmp.2005.11.006>
- Navarro, D. J., Myung, I. J., Pitt, M. A., & Kim, W. (2003). Global model analysis by landscaping. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (pp. 851–856). Austin, Texas: Cognitive Science Society.
- Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, 49(1), 47–84. <https://doi.org/10.1016/j.cogpsych.2003.11.001>
- Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In *Visual Information Processing: Proceedings of the Eighth Annual Carnegie Symposium on Cognition, Held at the Carnegie-Mellon University, Pittsburgh, Pennsylvania, May 19, 1972* (pp. 283–308). New York, London: Academic Press. <https://doi.org/10.1016/B978-0-12-170150-5.50012-3>
- Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, 10(8), 20, 1–19. <https://doi.org/10.1167/10.8.20>
- O'Keefe, J. (1990). A computational theory of the hippocampal cognitive map. In J. Storm Mathisen, J. Zimmer, & O. Ottersen (Eds.), *Understanding the brain through the hippocampus: The hippocampal region as a model for studying brain structure and function* (Vol. 83, pp. 301–312). Amsterdam, New York, Oxford: Elsevier. [https://doi.org/10.1016/S0079-6123\(08\)61258-3](https://doi.org/10.1016/S0079-6123(08)61258-3)
- O'Keefe, J. (1996). The spatial prepositions in English, vector grammar, and the cognitive map theory. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.), *Language and space* (pp. 277–316). Cambridge, Massachusetts; London, England: MIT Press.
- O'Keefe, J. (2003). Vector grammar, places, and the functional role of the spatial prepositions in English. In E. van der Zee & J. Slack (Eds.), *Representing direction in language and space* (pp. 69–85). Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199260195.003.0004>
- O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Oxford University Press. (available from <http://www.cognitivemap.net/>)
- O'Keefe, J., & Nadel, L. (1979). Précis of O'Keefe & Nadel's The hippocampus as a cognitive map. *Behavioral and Brain Sciences*, 2(4), 487–533. <https://doi.org/10.1017/S0140525X00063949>
- Olson, I. R., & Marshuetz, C. (2005). Remembering "what" brings along "where" in visual working memory. *Perception & Psychophysics*, 67(2), 185–194. <https://doi.org/10.3758/BF03206483>
- O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 46(3), 461–488. <https://doi.org/10.1037/h0084327>
- O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5), 939–1031. <https://doi.org/10.1017/S0140525X01000115>
- Palestro, J. J., Sederberg, P. B., Osth, A. F., van Zandt, T., & Turner, B. (2018). *Likelihood-free methods for cognitive science*. Cham, Switzerland: Springer. <https://doi.org/10.1007/978-3-319-72425-6>
- Pecher, D., & Zwaan, R. A. (Eds.). (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511499968>
- Pederson, E. (2003). How many reference frames? In C. Freksa, W. Brauer, C. Habel, & K. F. Wender (Eds.), *Spatial cognition III. Spatial cognition 2002*. (Vol. 2685, pp. 287–304). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-45004-1_17
- Peebles, D., & Cooper, R. P. (2015). Thirty years after Marr's Vision: Levels of analysis in cognitive science. *Topics in Cognitive Science*, 7(2), 187–190. <https://doi.org/10.1111/tops.12137>
- Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., & Spivey, M. J. (2013). Computational grounded cognition: A new alliance between grounded cognition and computational modeling. *Frontiers in Psychology*, 3, 612. <https://doi.org/10.3389/fpsyg.2012.00612>
- Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, 113(1), 57–83. <https://doi.org/10.1037/0033-295X.113.1.57>

- Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, 6(10), 421–425. [https://doi.org/10.1016/S1364-6613\(02\)01964-2](https://doi.org/10.1016/S1364-6613(02)01964-2)
- Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1), 3–25. <https://doi.org/10.1080/00335558008248231>
- Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, 109(2), 160–174. <https://doi.org/10.1037/0096-3445.109.2.160>
- Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, 32(1), 65–97. [https://doi.org/10.1016/0010-0277\(89\)90014-0](https://doi.org/10.1016/0010-0277(89)90014-0)
- Pylyshyn, Z. W. (1994). Some primitive mechanisms of spatial attention. *Cognition*, 50(1–3), 363–384. [https://doi.org/10.1016/0010-0277\(94\)90036-1](https://doi.org/10.1016/0010-0277(94)90036-1)
- Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends in Cognitive Sciences*, 4(5), 197–207. [https://doi.org/10.1016/S1364-6613\(00\)01477-7](https://doi.org/10.1016/S1364-6613(00)01477-7)
- Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80(1–2), 127–158. [https://doi.org/10.1016/S0010-0277\(00\)00156-6](https://doi.org/10.1016/S0010-0277(00)00156-6)
- Pylyshyn, Z. W. (2009). Perception, representation, and the world: The FINST that binds. In D. Dedrick & L. Trick (Eds.), *Computation, cognition, and Pylyshyn* (pp. 3–48). Cambridge, Massachusetts: MIT Press.
- Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, 3(3), 179–197. <https://doi.org/10.1163/156856888X00122>
- R Core Team. (2016). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. (<https://www.R-project.org/>)
- Raichle, M. E. (2010). Two views of brain function. *Trends in Cognitive Sciences*, 14(4), 180–190. <https://doi.org/10.1016/j.tics.2010.01.008>
- Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. Cambridge, Massachusetts: MIT Press.
- Regier, T. (1997). Constraints on the learning of spatial terms: A computational investigation. In R. L. Goldstone, D. L. Medin, & P. G. Schyns (Eds.), *Perceptual Learning* (Vol. 36, pp. 171–219). San Diego, California: Academic Press. [https://doi.org/10.1016/S0079-7421\(08\)60284-4](https://doi.org/10.1016/S0079-7421(08)60284-4)
- Regier, T., & Carlson, L. A. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2), 273–298. <https://doi.org/10.1037/0096-3445.130.2.273>
- Regier, T., & Xu, Y. (2017). The Sapir-Whorf hypothesis and inference under uncertainty. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(6), e1440. <https://doi.org/10.1002/wcs.1440>
- Richardson, D. C., Spivey, M. J., Barsalou, L. W., & McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. *Cognitive Science*, 27(5), 767–780. https://doi.org/10.1207/s15516709cog2705_4
- Richardson, D. C., Spivey, M. J., Edelman, S., & Naples, A. D. (2001). “Language is spatial”: Experimental evidence for image schemas of concrete and abstract verbs. In J. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society* (pp. 873–878). Austin, Texas: Cognitive Science Society.
- Richter, M. (2018). *A neural dynamic model for the perceptual grounding of spatial and movement relations* (Unpublished doctoral dissertation). Ruhr-Universität Bochum, Institut für Neuroinformatik. (URN: urn:nbn:de:hbz:294-60740)
- 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, Texas: Cognitive Science Society.
- Richter, M., Lins, J., & Schöner, G. (2016). A neural dynamic model parses object-oriented actions. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*. (pp. 1931–1936). Austin, Texas: 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(1), 35–47. <https://doi.org/10.1111/tops.12240>
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2), 358–367. <https://doi.org/10.1037/0033-295X.107.2.358>
- Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4), 515–530. <https://doi.org/10.1111/j.1467-9868.2005.00513.x>
- Rosielle, L. J., Crabb, B. T., & Cooper, E. E. (2002). Attentional coding of categorical relations in scene perception: Evidence from the flicker paradigm. *Psychonomic Bulletin & Review*, 9(2), 319–326. <https://doi.org/10.3758/BF03196288>
- Roth, J. C., & Franconeri, S. L. (2012). Asymmetric coding of categorical spatial relations in both language and vision. *Frontiers in Psychology*, 3, 464. <https://doi.org/10.3389/fpsyg.2012.00464>
- Roy, D., & Mukherjee, N. (2005). Towards situated speech understanding: Visual context priming of language models. *Computer Speech & Language*, 19(2), 227–248. <https://doi.org/10.1016/j.csl.2004.08.003>
- Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing: Ex-*

- plorations in the microstructure of cognition: Foundations* (Vol. 1). Cambridge, Massachusetts: MIT Press.
- Samuelson, L. K., Smith, L. B., Perry, L. K., & Spencer, J. P. (2011). Grounding word learning in space. *PLOS ONE*, *6*(12), e28095. <https://doi.org/10.1371/journal.pone.0028095>
- Sanderson, C., & Curtin, R. (2016). Armadillo: a template-based C++ library for linear algebra. *Journal of Open Source Software*, *1*(2), 26. <https://doi.org/10.21105/joss.00026>
- Savelli, F., & Knierim, J. J. (2018). AI mimics brain codes for navigation. *Nature*, *557*, 313–314. <https://doi.org/10.1038/d41586-018-04992-7>
- Scheider, S., Hahn, J., Weiser, P., & Kuhn, W. (2018). Computing with cognitive spatial frames of reference in GIS. *Transactions in GIS*, *22*(5), 1083–1104. <https://doi.org/10.1111/tgis.12318>
- Schöner, G. (2008). Dynamical systems approaches to cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 101–126). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511816772.007>
- Schultheis, H., & Carlson, L. A. (2017). Mechanisms of reference frame selection in spatial term use: Computational and empirical studies. *Cognitive Science*, *41*(2), 276–325. <https://doi.org/10.1111/cogs.12327>
- Schultheis, H., & Carlson, L. A. (2018). Inter-process relations in spatial language: Feedback and graded compatibility. *Cognition*, *176*, 140–158. <https://doi.org/10.1016/j.cognition.2018.02.020>
- Schultheis, H., Singhaniya, A., & Chaplot, D. S. (2013). Comparing model comparison methods. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 1294–1299). Austin, Texas: Cognitive Science Society.
- Schutte, A. R., & Spencer, J. P. (2009). Tests of the dynamic field theory and the spatial precision hypothesis: Capturing a qualitative developmental transition in spatial working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1698–1725. <https://doi.org/10.1037/a0015794>
- Schöner, G., Spencer, J., & the DFT Research Group. (2016). *Dynamic thinking: A primer on dynamic field theory*. Oxford: Oxford University Press.
- Shiffrin, R. M. (2010). Perspectives on modeling in cognitive science. *Topics in Cognitive Science*, *2*(4), 736–750. <https://doi.org/10.1111/j.1756-8765.2010.01092.x>
- Shusterman, A., & Li, P. (2016). Frames of reference in spatial language acquisition. *Cognitive Psychology*, *88*, 115–161. <https://doi.org/10.1016/j.cogpsych.2016.06.001>
- Smith, L. B., Maouene, J., & Hidaka, S. (2007). The body and children's word learning. In J. M. Plumert & J. P. Spencer (Eds.), *The emerging spatial mind* (pp. 168–192). Oxford: Oxford University Press.
- Spivey, M. J., Richardson, D. C., & Fitneva, S. A. (2004). Thinking outside the brain: Spatial indices to visual and linguistic information. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 161–189). New York, Hove: Psychology Press.
- Stan Development Team. (2016). RStan: The R interface to Stan [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=rstan> (R package version 2.14.1)
- Stocker, K., & Laeng, B. (2017). Analog and digital windowing of attention in language, visual perception, and the brain. *Cognitive Semantics*, *3*(2), 158–181. <https://doi.org/10.1163/23526416-00302002>
- Sun, R. (Ed.). (2008). *The Cambridge handbook of computational psychology*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511816772>
- Sun, R. (2009). Theoretical status of computational cognitive modeling. *Cognitive Systems Research*, *10*(2), 124–140. <https://doi.org/10.1016/j.cogsys.2008.07.002>
- Talmy, L. (1983). How language structures space. In H. L. Pick & L. P. Acredolo (Eds.), *Spatial orientation* (pp. 225–282). Boston, Massachusetts: Springer. https://doi.org/10.1007/978-1-4615-9325-6_11
- Talmy, L. (2000). *Towards a cognitive semantics* (Vol. I: Concept Structuring Systems). Cambridge, Massachusetts; London, England: MIT Press.
- Tanenhaus, M. K., Spivey Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. <https://doi.org/10.1126/science.7777863>
- Tenbrink, T., & Kuhn, W. (2011). A model of spatial reference frames in language. In M. Egenhofer, N. Giudice, R. Moratz, & M. Worboys (Eds.), *Spatial information theory, COSIT 2011* (Vol. 6899, pp. 371–390). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-23196-4_20
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. <https://doi.org/10.1037/h0061626>
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. [https://doi.org/10.1016/0010-0285\(80\)90005-5](https://doi.org/10.1016/0010-0285(80)90005-5)
- Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin & Review*, *21*(2), 227–250. <https://doi.org/10.3758/s13423-013-0530-0>
- Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, *56*(2), 69–85. <https://doi.org/10.1016/j.jmp.2012.02.005>
- Ullman, S. (1984). Visual routines. *Cognition*, *18*(1–3), 97–159. [https://doi.org/10.1016/0010-0277\(84\)90023-4](https://doi.org/10.1016/0010-0277(84)90023-4)
- Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, compet-

- ing accumulator model. *Psychological Review*, 108(3), 550–592. <https://doi.org/10.1037//0033-295X.108.3.550>
- Vandeloise, C. (1991). *Spatial prepositions: A case study from French*. Chicago, London: University of Chicago Press. (translated by Anna R. K. Bosch)
- van der Ham, I. J., & Postma, A. (2010). Lateralization of spatial categories: A comparison of verbal and visuospatial categorical relations. *Memory & Cognition*, 38(5), 582–590. <https://doi.org/10.3758/MC.38.5.582>
- van der Ham, I. J., Postma, A., & Laeng, B. (2014). Lateralized perception: The role of attention in spatial relation processing. *Neuroscience & Biobehavioral Reviews*, 45, 142–148. <https://doi.org/10.1016/j.neubiorev.2014.05.006>
- van Oeffelen, M. P., & Vos, P. G. (1982). Configurational effects on the enumeration of dots: Counting by groups. *Memory & Cognition*, 10(4), 396–404. <https://doi.org/10.3758/BF03202432>
- van Oeffelen, M. P., & Vos, P. G. (1983). An algorithm for pattern description on the level of relative proximity. *Pattern Recognition*, 16(3), 341–348. [https://doi.org/10.1016/0031-3203\(83\)90040-7](https://doi.org/10.1016/0031-3203(83)90040-7)
- Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, 7(3), 424–465. <https://doi.org/10.3758/BF03214357>
- Vehtari, A., Gelman, A., & Gabry, J. (2016). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=loo> (R package version 1.0.0.)
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- Veksler, V. D., Myers, C. W., & Gluck, K. A. (2015). Model flexibility analysis. *Psychological Review*, 122(4), 755–769. <https://doi.org/10.1037/a0039657>
- Vishwanath, D., & Kowler, E. (2003). Localization of shapes: Eye movements and perception compared. *Vision Research*, 43(15), 1637–1653. [https://doi.org/10.1016/S0042-6989\(03\)00168-8](https://doi.org/10.1016/S0042-6989(03)00168-8)
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, 14(5), 779–804. <https://doi.org/10.3758/BF03194105>
- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25(1), 35–57. <https://doi.org/10.3758/s13423-017-1343-3>
- Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, 48(1), 28–50. <https://doi.org/10.1016/j.jmp.2003.11.004>
- Wilson, H. R., & Kim, J. (1994). Perceived motion in the vector sum direction. *Vision Research*, 34(14), 1835–1842. [https://doi.org/10.1016/0042-6989\(94\)90308-5](https://doi.org/10.1016/0042-6989(94)90308-5)
- Yantis, S. (2000). Goal-directed and stimulus-driven determinants of attentional control. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (pp. 73–103). Cambridge, Massachusetts; London, England: MIT Press.
- Yuan, L., Uttal, D., & Franconeri, S. (2016). Are categorical spatial relations encoded by shifting visual attention between objects? *PLOS ONE*, 11(10), e0163141. <https://doi.org/10.1371/journal.pone.0163141>
- Zwarts, J. (1997). Vectors as relative positions: A compositional semantics of modified PPs. *Journal of Semantics*, 14(1), 57–86. <https://doi.org/10.1093/jos/14.1.57>
- Zwarts, J. (2017). Spatial semantics: Modeling the meaning of prepositions. *Language and Linguistics Compass*, 11(5), e12241. <https://doi.org/10.1111/lnc3.12241>
- Zwarts, J., & Gärdenfors, P. (2016). Locative and directional prepositions in conceptual spaces: The role of polar convexity. *Journal of Logic, Language and Information*, 25(1), 109–138. <https://doi.org/10.1007/s10849-015-9224-5>