Quantifying quantifier representations:

Experimental studies, computational modeling, and individual differences

Sonia Ramotowska

Quantifying quantifier representations:

Experimental studies, computational modeling, and individual differences

ILLC Dissertation Series DS-2022-03

Institute for Logic, Language and Computation

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

Cover design by Sonia Ramotowska and Ipskamp Printing.
Printed and bound by Ipskamp Printing.

ISBN: 978-94-6421-703-2

Quantifying quantifier representations:
Experimental studies, computational modeling, and individual differences

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op maandag 16 mei 2022, te 14.00 uur

door Sonia Ramotowska
geboren te Warschau

Promotiecommissie

Promotor:	prof. dr. S.J.L. Smets	Universiteit van Amsterdam
Copromotores:	dr. J.K. Szymanik	Universiteit van Amsterdam
	dr. L. van Maanen	Universiteit Utrecht
Overige leden:	prof. dr. J.E. Rispens	Universiteit van Amsterdam
	prof. dr. F.X. Alario	CNRS
	prof. dr. rer. nat. S. Heim	Institute of Neuroscience and Medicine Forschungszentrum Jülich
	prof. dr. B. Kaup	Eberhard Karls Universität Tübingen
	dr. M.D. Aloni	Universiteit van Amsterdam
	dr. S. Pezzelle	Universiteit van Amsterdam

Faculteit der Geesteswetenschappen

The research for this doctoral thesis received financial assistance from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement n.STG 716230 CoSaQ.

To my parents and my brother.

Contents

Acknowledgments **xiii**

Author Contributions **xvii**

1 Introduction **1**
1.1 Quantifiers in logic and formal linguistics 2
1.1.1 Basic formal notions 2
1.1.2 The logical view on quantifiers 5
1.2 Toward cognitive perspective on quantifiers 9
1.2.1 Current computational models of quantifiers 10
1.2.2 New modeling approach 11
1.3 Thesis overview 12
1.3.1 How are quantifiers ordered? 13
1.3.2 How are quantifiers represented and processed? 13
1.3.3 Why are some quantifiers more difficult to process than others? 13
1.3.4 How do formal properties of quantifiers link to human cognitive abilities? 14

2 *Most* quantifiers have *many* meanings **15**
2.1 Introduction 15
2.1.1 How many subgroups of participants with different meanings can we identify? 17
2.1.2 How are the meanings of quantifiers interrelated at the subject level? 18
2.1.3 How are the parameters of our model interrelated? 19
2.1.4 Current study 20
2.2 Method 22
2.2.1 Participants 22

2.2.2 Experimental Design and Procedure . . . 23
2.2.3 Data pre-processing . . . 24
2.2.4 Computational Model . . . 24
2.2.5 Cluster analysis . . . 26
2.2.6 Linear Discriminant Analysis . . . 26
2.3 Results . . . 26
2.3.1 Estimated parameters . . . 26
2.3.2 Cluster analysis results . . . 30
2.3.3 Linear Discriminant Analysis results . . . 36
2.4 Discussion . . . 38
2.4.1 Order of quantifiers on the mental line . . . 38
2.4.2 Relationship between model parameters . . . 40
2.4.3 Sources of individual differences . . . 41
2.4.4 Conclusions . . . 42

3 Uncovering the structure of semantic representations using a computational model of decision-making 43
3.1 Introduction . . . 44
3.1.1 Semantic representations: the case of *most* and *more than half* . . . 45
3.1.2 Modeling, experiments, and predictions . . . 48
3.2 Experiment 1 . . . 52
3.2.1 Methods . . . 52
3.2.2 Computational modeling of the verification process . . . 54
3.2.3 Results . . . 56
3.3 Experiment 2 . . . 64
3.3.1 Methods . . . 64
3.3.2 Computational model . . . 65
3.3.3 Results . . . 66
3.4 Discussion . . . 71

4 Representational complexity and pragmatics cause the monotonicity effect 75
4.1 Background and goals . . . 75
4.2 Competing theoretical proposals . . . 76
4.3 Main ingredients of the DDM . . . 77
4.4 Link to theoretical proposals . . . 78
4.5 Methods . . . 78
4.5.1 Participants . . . 79
4.5.2 Design, materials, and procedures . . . 79
4.5.3 Exclusion criteria . . . 80
4.5.4 Regression analyses and modeling strategy . . . 81
4.6 Results . . . 81

4.6.1 Regression Analyses . . . 81
4.6.2 DDM analyses . . . 83
4.7 Discussion . . . 85

5 Discovering stages of processing in quantified sentences 87
5.1 Introduction . . . 87
5.1.1 The two-step models . . . 89
5.1.2 Electroencephalography evidence for two-step model . . . 92
5.1.3 Hidden semi-Markov Model Multivariate Pattern Analysis (HsMM-MVPA) . . . 94
5.2 Methods . . . 97
5.2.1 Participants . . . 98
5.2.2 Experimental design . . . 98
5.2.3 Procedure . . . 99
5.2.4 EEG recording . . . 99
5.2.5 Choice of analysis time windows . . . 99
5.2.6 EEG data preprocessing . . . 101
5.2.7 Statistical analysis of reaction times . . . 101
5.2.8 ERP analyses . . . 102
5.2.9 HsMM-MVPA . . . 104
5.2.10 Statistical analysis of stage durations . . . 105
5.3 Results . . . 105
5.3.1 Reaction time analysis . . . 105
5.3.2 ERP analyses . . . 106
5.3.3 HsMM-MVPA . . . 111
5.3.4 HsMM-MVPA mapped models . . . 116
5.3.5 Do stages predict the length of reaction times? . . . 118
5.3.6 Stage durations analysis . . . 119
5.4 General Discussion . . . 121
5.4.1 After the quantifier onset . . . 121
5.4.2 After the adjective onset . . . 125
5.4.3 Alternative explanation of the polarity effect . . . 129
5.4.4 Methodological implications . . . 134
5.4.5 Conclusions . . . 138

6 Does ease of learning explain quantifier universals? 139
6.1 Introduction . . . 139
6.1.1 Quantifiers . . . 141
6.1.2 Monotonicity and convexity (connectedness) . . . 142
6.1.3 Quantity . . . 144
6.1.4 Conservativity . . . 144
6.1.5 Current experiment - predictions . . . 146
6.2 Methods . . . 147

6.2.1 Participants . . . 147
6.2.2 Materials and design . . . 147
6.2.3 Procedure . . . 150
6.2.4 Statistical analysis . . . 150
6.2.5 Frequency analysis of quantifiers . . . 151
6.3 Results . . . 151
6.3.1 Monotonicity . . . 153
6.3.2 Convexity . . . 154
6.3.3 Quantity . . . 154
6.3.4 Conservativity . . . 155
6.4 Discussion . . . 155
6.4.1 Monotonicity and Convexity . . . 156
6.4.2 Quantity . . . 157
6.4.3 Conservativity . . . 158
6.4.4 Methodological remarks . . . 158
6.4.5 Potential confounds . . . 160
6.4.6 Minimal pair methodology limitations . . . 163
6.4.7 Conclusions . . . 163

7 Conclusions **165**
7.1 Summary of the main findings . . . 165
7.2 Cognitive model . . . 167
7.2.1 Probabilistic semantics of quantifiers . . . 168
7.2.2 Fuzzy quantifiers . . . 169
7.2.3 Modeling of thresholds and vagueness . . . 170
7.3 Polarity effect . . . 171
7.4 Relationship between computational models . . . 172
7.5 Coda . . . 174

A Appendix to Chapter 2 **175**

B Appendix to Chapter 3 **177**
B.1 Additional analyses . . . 177
B.1.1 DV hypothesis testing - regression model comparison . . . 177
B.1.2 Effect of threshold on reaction time – regression model comparison . . . 177
B.1.3 Additional findings . . . 178
B.2 Replication experiment . . . 179
B.2.1 Methods . . . 179
B.2.2 Results . . . 180

C Appendix to Chapter 5 **187**
C.1 Reaction time analysis without timeout 187
C.2 ERP analyses 188
C.2.1 ERPs after the quantifier onset - ANOVA 188
C.2.2 ERPs after the adjective onset - ANOVA 189
C.3 HsMM-MVPA - 500 ms time window after the quantifier onset . . 190
C.4 HsMM-MVPA - mean log-likelihood 192
C.4.1 After the quantifier onset 192
C.4.2 After the adjective onset 194
C.5 Model comparison 195
C.5.1 After the quantifier onset 195
C.5.2 After the adjective onset 197
C.6 HsMM-MVPA of long sentences 200
C.6.1 Stage 7 204
C.6.2 Stage 9 204
C.7 Stage 11 analysis without timeout 204

D Appendix to Chapter 6 **205**
D.1 Bayesian logistic regression model diagnostics 205
D.2 Graphical representation of Bayesian logistic regression model estimates 206
D.3 Models of quantifier frequency 207

References **209**

Samenvatting **229**

Abstract **231**

Acknowledgments

Carrying out interdisciplinary research is enormously challenging. I was extremely lucky to have supervisors that helped me to approach my research topic from two completely different, yet complementary perspectives. My research, but also I, as a scientist, benefited a lot from their combined skills and expertise.

I came to the CoSaQ project without extensive training in formal semantics to go off the deep end and carry out research on quantifiers. I am grateful to Jakub for helping me on the path towards experimental semantics. He was always available to answer my question or discuss progress in a research project. Many times, I could just drop by his office or have a chat online about the ongoing project. I especially value his attention to detail and at the same time the ability to keep the big picture in mind. I am also grateful for his patience and feedback on my rough drafts of the papers as well as thesis introduction and conclusions. In addition to his support in carrying out my research project, I thank him for his advice on my career development and his openness to approach him with (even small) practical matters.

I am grateful to Leendert for bringing the innovative, methodological approach into my research. I admire his passion for computational modeling and all the support I received once I started exploring these completely new methods. I am deeply grateful for his insightful comments on the manuscript and paper drafts and for helping me to bridge cognitive science to semantics. Additionally, I thank Leendert for the translation of the Samenvatting of this thesis.

I am honored to have my thesis evaluated by the excellent experts in psycholinguistics, semantics, and modeling: prof. dr. Rispens, prof. dr. Alario, prof. dr. Heim, prof, dr. Kaup, dr. Aloni and dr. Pezzelle. I would like to thank the members of the Doctorate Committee for agreeing to this task and the time they spend reading and assessing my dissertation. I would like to extend my gratitude to my promotor Sonja Smets for ensuring that my research is on a good track.

Interdisciplinary research is not possible without team working and collabora-

tion. I would like to thank the whole ILLC community for creating a collegial and inspiring environment. In particular, I am very deeply grateful to the members of the CoSaQ research group, Shane, Milica, and Fausto, for creating a supportive research atmosphere and helpful feedback at all stages of the research project.

I would like to express my deepest thanks to all my collaborators for their unique input in my research. Firstly, I thank Shane for helping me to make the first steps in the javascript programming of my experiments. I have never thought that I would find programming as much fun. I would like to thank Julia for working on developing a model to test individual differences in quantifier representations. I would like to thank Fabian for his attention to detail, precision, and clarity in formulating research questions. It was also a greater pleasure to collaborate with him on two research projects and present our results at the CUNY conference. I am very grateful to Petra and Fabian for sharing the EEG data from their experiment. Without their help, the project would most likely not be possible to accomplish in times of a global pandemic. I thank Kim for guiding me through the pre-processing of the EEG data. Many thanks go to Hermine for sharing her scripts with me, discussion of the HsMM-MVPA results, and feedback on the manuscript draft. Her insightful comments certainly contributed to a better quality of this project.

I would like to thank the organizers of the Meaning, Logic, and Cognition seminar at ILLC (Milica and Dean) for giving me the chance to present my work in progress and all attendees for their feedback. I am especially grateful to members of the CoSaQ reading group (Shane, Fausto, Milica, Saskia, Heming, and Master of Logic students: Lorenzo, Simone, and Terence) for their comments and suggestions on the pilot for the learnability experiment. I also much appreciate the opportunity to present the results of my EEG analysis at many seminars venues: at the University of Tübingen, Osnabrück, Groningen, Tilburg, and Aix-Marseille Université, and the feedback I received. I would like to thank especially warmly the members of the Cognitive Modeling Group at the University of Groningen for the engaging and fruitful discussion.

I thank warmly my office mates Natalia, Klaas, and Rosa. I very much enjoyed sharing the office with you before the pandemic time. In particular, I thank Natalia for the welcoming atmosphere in our office on the first day of my work. My first months in Amsterdam would be certainly more stressful and lonelier without her company.

I also want to thank my dear friend Tereska. I am happy to be her friend since primary school. For all these years, and in particular the last four, I could share with her all my ups and downs. I want to thank Aaron for inspiring me to pursue a scientific career and start my PhD. I also thank him for pushing me into jogging. As much as it was painful for me at the beginning, it has become my source of everyday energy and best thoughts organizer. The celebration of my 25th birthday was quite unusual for me. I moved to Amsterdam that day and started my 4 years adventure. I would like to thank Eftychia for hosting me when

I arrived in Amsterdam and for making me feel much less lonely (and of course for remembering about the birthday cake). An especially warm thank goes to Erwin. During the last months of my PhD, I would not maintain that much peace of mind without his companionship. Thanks to Erwin I discovered the wonderful taste of Indonesian and Surinamese cuisine and a passion for harvesting even in the most Dutch weather.

Moving abroad to continue my academic career would not be easy for me without the loving support from my family. Even kilometers away from home, I always felt close to my mother, father, and brother. My parents have always been supporting for me in developing my passion for science. They had deep confidence and faith in my abilities even when I had many doubts.

Amsterdam Sonia Ramotowska
March, 2022.

Author Contributions

The thesis consists of 7 chapters. **Chapter 1** (Introduction) and **Chapter 7** (Conclusions) were written by Sonia Ramotowska. Jakub Szymanik and Leendert van Maanen provided feedback to these chapters.

Chapters 2 to 6 correspond to the manuscripts already published, in revision or in preparation. The contribution to each chapter is defined according to CRediT.

Chapter 2. Most quantifiers have many meanings

Manuscript: Ramotowska, S., Haaf, J., van Maanen, L., Szymanik, J. Most quantifiers have many meanings

Conceptualization – SR, JH, LvM, JS; Data curation – SR; Formal analysis – SR, JH; Funding acquisition - JS; Investigation – SR; Methodology – SR, JH, LvM, JS; Software – SR, JH; Supervision – LvM, JS; Visualization – SR, JH; Writing – original draft – SR; Writing – review & editing – SR, JH, LvM, JS

Chapter 3. Uncovering the structure of semantic representations using a computational model of decision-making

Manuscript: Ramotowska S., Steinert-Threlkeld S., van Maanen L., Szymanik J. (2021). Uncovering the structure of semantic representations using a computational model of decision-making, *in revision*.

Conceptualization – SR, SST, LvM, JS; Data curation – SR; Formal analysis – SR; Funding acquisition - JS; Investigation – SR, SST; Methodology – SR, SST, LvM, JS; Software – SR, SST, LvM; Supervision – LvM, JS; Validation – SR; Visualization – SR; Writing – original draft – SR; Writing – review & editing – SR, SST, LvM, JS.

The chapter is also a version of an unpublished manuscript:

Ramotowska S., Steinert-Threlkeld S., van Maanen L., Szymanik J. (2020). Individual differences in semantic representations: The case of *most* and *more*

than half (available online on Semantics Archive).

Chapter 4. Representational complexity and pragmatics cause the monotonicity effect

This chapter includes the publication:

Manuscript: Schlotterbeck F., Ramotowska S., van Maanen L., Szymanik J. (2020). Representational complexity and pragmatics cause the monotonicity effect. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3397-3403). Cognitive Science Society (shared first authorship).

Conceptualization – SR, FS, LvM, JS; Data curation – SR, FS; Formal analysis – SR, FS; Funding acquisition - JS, FS; Investigation – SR, FS; Methodology – SR, FS, LvM, JS; Software – SR, FS, LvM; Supervision – LvM, JS; Visualization – SR, FS; Writing – original draft – SR, FS; Writing – review & editing – SR, FS, LvM, JS

SR made minor editorial changes in this chapter compared to the published manuscript.

Chapter 5. Discovering stages of processing in quantified sentences

Manuscript: Ramotowska, S., Archambeau, K., Augurzky, P., Schlotterbeck, F., Berberyan, H., van Maanen, L., Szymanik, J. Discovering stages of processing in quantified sentences

Conceptualization – SR, PA, FS, LvM, JS; Data curation – SR, FS; Formal analysis – SR, KA, HB; Funding acquisition - PA, JS; Investigation – PA, SR; Methodology – SR, PA, FS, HB, LvM, JS; Software – SR, KA, HB; Supervision – LvM, JS; Visualization – SR, HB; Writing – original draft – SR; Writing – review & editing – SR, KA, PA, FS, HB, LvM, JS

The analyses presented in this chapter were conducted on data collected and published previously:

Augurzky, P., Schlotterbeck, F., Ulrich, R. (2020). Most (but not all) quantifiers are interpreted immediately in visual context. *Language, Cognition and Neuroscience, 35 (9)*, 1203–1222.

The data was shared by FS and PA. Their research was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 75650358 - SFB 833, project B1.

Chapter 6. Does ease of learning explain quantifier universals?

Manuscript: Ramotowska, S., van Maanen, L., Szymanik, J. Does ease of learning explain quantifier universals?

Conceptualization – SR, LvM, JS; Data curation – SR; Formal analysis –

SR; Funding acquisition - JS; Investigation – SR; Methodology – SR, LvM, JS; Software – SR; Supervision – LvM, JS; Visualization – SR; Writing – original draft – SR; Writing – review & editing – SR, LvM, JS

Chapter 1

Introduction

All natural languages have some ways to convey information about quantities. For example, in English, quantitative information can be expressed with numbers:

1.0.1. Example.

1. 60% of the tourists coming to Amsterdam take a boat tour.
2. I ate 100 pepernoten last week.

or by using quantifiers

3. *More than half* of the tourists coming to Amsterdam take a boat tour.
4. I ate *many* pepernoten last week.

Intuitively, quantifiers are natural language expressions that communicate quantitative information. Like other natural languages, English is also reach in quantified expressions, for example: *some*, *many*, *more than half*, *few*, *most*, *none*, *all*, and *at least 90*. This broad class of expressions varies in properties. Some quantifiers like *all* or *none* have very specific meanings that could be mapped on exact proportion or number (e.g., *none* means zero and *all* means 100%). Other quantifiers are less specific, yet can still be mapped on numerical information (e.g., *more than half*). The reader would probably agree that 60% of tourists consist of *more than half* of the tourists. There are also very vague quantifiers (e.g., *few*, *many*) that can have various meanings for different individuals and in different contexts. For example, the meaning of the sentence "I ate *many* pepernoten last week" may depend on the reader's affinity with Dutch sweets.

Quantifiers commonly used in everyday language have well-studied mathematical properties (Mostowski, 1957; Lindström, 1966). As mathematical objects, they have been investigated in formal linguistics (Montague, 1970; Barwise & Cooper, 1981) for around 50 years already. More recently, quantifiers have inspired an explosive amount of studies in cognitive science, psychology, and

experimental semantics[1]. The experimental data on quantifiers has increased the demand for applying computational modeling (e.g., Schöller & Franke, 2016; Schlotterbeck, 2017; van Tiel et al., 2021; Carcassi & Szymanik, 2021). In the introduction to this thesis, I will provide the most important formal notions and definitions of quantifiers. Next, I will contrast two perspectives on quantifiers: logical and cognitive. The *logical perspective* on quantifiers was taken in the majority of experimental studies. I will show the limitations of this view and argue in favor of the *cognitive perspective*. Finally, I will show how computational modeling can advance experimental studies on quantifiers and provide a brief overview of the thesis content.

1.1 Quantifiers in logic and formal linguistics

1.1.1 Basic formal notions

Andrzej Mostowski (1957) is a father of the modern mathematical Generalized Quantifiers Theory. *Generalized quantifiers* refer to the generalizations of two logical quantifiers: existential $\exists$ and universal $\forall$. By using the model theory, Mostowski provided a mathematical foundation of quantification. He defined the notion of generalized quantifiers of the type $< 1 >$. Intuitively, type $< 1 >$ quantifiers define the properties of sets, for example, the universal quantifier $\forall$ denotes the property of *being identical to the universe*, and existential quantifier $\exists$ the property of *being not-empty*.

Most of the quantifiers in natural language are of type $< 1, 1 >$. While quantifiers of type $< 1 >$ denote the properties of sets, quantifiers of type $< 1, 1 >$ denote *relationships* between sets. Therefore, they are also called *quantirelations* (Peters & Westerståhl, 2008). We owe the extension of the generalized quantifier definition to arbitrary type to Lindström (1966). While Lindström's (1966) definition is applied mostly in logic (cf. Szymanik, 2016), in linguistics the generalized quantifiers (of arbitrary type) are often defined as a *relationship between relationships*.

1.1.1. Definition. A generalized quantifier (Q) of arbitrary type $< n_1, ..., n_k >$ is a function that assigns to every universe of discourse (M) a k-ary relation Q_M

[1]E.g., Hackl (2009); Pietroski, Lidz, Hunter, and Halberda (2009); Pietroski, Lidz, Hunter, Odic, and Halberda (2011); Szymanik and Zajenkowski (2013); Zajenkowski and Szymanik (2013); Zajenkowski, Szymanik, and Garraffa (2014); Deschamps, Agmon, Loewenstein, and Grodzinsky (2015); Heim et al. (2015); Shikhare, Heim, Klein, Huber, and Willmes (2015); Schöller and Franke (2016); Schlotterbeck (2017); Talmina, Kochari, and Szymanik (2017); Pezzelle, Bernardi, and Piazza (2018); Agmon, Loewenstein, and Grodzinsky (2019); Ramotowska, Steinert-Threlkeld, van Maanen, and Szymanik (2020b); Denić and Szymanik (2020); Heim, Peiseler, and Bekemeier (2020); Carcassi and Szymanik (2021); van Tiel, Franke, and Sauerland (2021).

between relationships on M such that if $(R_1, ..., R_k) \in Q_M$, then R_i is an n_i-ary relation on M, for $i = 1,..., k$ (see Szymanik, 2016, pp. 26).

In other words, the quantifier of type $< 1, 1 >$ is a higher-order function that assigns the truth value to the relationship between two subsets of the universe of discourse. In natural language, *some*, *most*, and *many* are the examples of type $< 1, 1 >$ quantifiers. Given a universe of discourse M and two subsets of M, A and B, the quantifiers *some*, *most*, and *many* are true if and only if the following truth conditions are met:

1.1.2. Example.

1. $some_M(A, B) = 1$ iff $|A \cap B| \neq 0$
2. $most_M(A, B) = 1$ iff $|A \cap B| > |A - B|$
3. $many_M(A, B) = 1$ iff $|A \cap B| > d$, where d means contextual threshold.

The important addition to the definition of quantifiers is that their meanings do not depend on the specific properties of A and B subsets. This property is called *topic neutrality*. All that matters for the quantifier to be true is the size of relevant sets, but not the specific properties of the elements of those sets. Formally, the topic neutrality is referred to as *isomorphism*[2].

In parallel to the development of the Generalized Quantifier Theory in logic, quantifiers also gained attention in formal linguistics. Montague (1970) and Barwise and Cooper (1981) are the founding fathers of the linguistic theory of quantifiers. Their main contribution was to link the logical syntax of generalized quantifiers with natural language syntax. The most important class of quantifiers of type $< 1, 1 >$ in natural language are called determiners. These are expressions such as: *some*, *the*, *most*, *all but ten*, and *not every*. In combination with the nouns, the determiners *restrict* the sentence to the subset of the universe of discourse. Together, the determiner and the noun constitute the noun phrase (NP, e.g., *some cats*, *the boy*, *most dogs*, *all but ten girls*, *not every day*). Once the NP is combined with a verb phrase (VP, the *scope* of the sentence, e.g., *chirp*, *sleeps*, *bark*, *went to the party*, *is sunny*) the *quantified sentences* is formulated:

1.1.3. Example.

1. *Some cats chirp.*
2. *The boy sleeps.*
3. *Most dogs bark.*

[2]I will discuss the isomorphism property in detail in Chapter 6.

4. *All but ten girls went to the party.*

5. *Not every day is sunny.*

Using the quantifier's truth condition, we can define the truth condition for a quantified sentence in natural language. For example, we can define the truth condition for the sentence "*Most* dogs bark" based on the definition of *most* in EXAMPLE 1.1.2:

1.1.4. EXAMPLE. $|[\![DOG]\!] \cap [\![BARK]\!]| > |[\![DOG]\!] - [\![BARK]\!]|$

Another contribution of Barwise and Cooper (1981) to formal semantics was to discuss and define the main, formal properties of natural language quantifiers. They introduced the notion of the *live-on property*, nowadays called *conservativity*[3]. The quantifier Q of type $< 1, 1 >$ is conservative if:

1.1.5. DEFINITION. $Q(A, B)$ iff $Q(A, A \cap B)$.

Conservativity means that the meaning of the quantifier of type $< 1, 1 >$ refers only to the restrictor and intersection of the sets[4].

Moreover, Barwise and Cooper (1981) analyzed the *monotonicity* property of quantifiers. Monotonicity is one of the basic properties in natural language that refers to entailment patterns. A quantifier is *monotone* if it is either *upward* or *downward* monotone (Barwise & Cooper, 1981)[5]. Moreover, a quantifier can be upward or downward monotone on the left, right, or both arguments. Monotonicity on the left argument concerns the restrictor, and on the right argument the scope. Monotonicity on the left argument determines whether the truth value of the sentence changes when we add or remove elements in the quantifier's restrictor. Monotonicity on the right argument, in turn, determines the truth value of the quantifier if the number of elements changes in the scope. Let us have a closer look at two intuitive examples. The quantifier *more than half* is upward monotone on the right argument because when we extend its scope, the truth value of the sentence remains unchanged. For *more than half* the sentence (2) in EXAMPLE 1.1.6 entails the sentence (1). If *more than half* of the students passed the difficult exam, then it is also true that *more than half* of the students passed the exam, because the scope *exam* is more general than *difficult exam*. The opposite entailment holds for *fewer than half*. Narrowing of the scope preserves the truth value of the downward monotone quantifiers on the right argument.

[3]The term *conservativity* was introduced by Keenan and Stavi (1986).

[4]I will discuss the conservativity property in detail in Chapter 6.

[5]I will discuss the monotonicity and the related polarity property in details in Chapters 4, 5 and 6.

1.1.6. Example.

1. *More than half / Fewer than half* of the students passed the exam.
2. *More than half / Fewer than half* of the students passed the difficult exam.

Formally, the quantifier Q of type $< 1, 1 >$ is

1.1.7. Definition. upward monotone on the right argument when if $Q(A, B)$ and $B \subseteq B'$, then $Q(A, B')$

1.1.8. Definition. downward monotone on the right argument when if $Q(A, B)$ and $B' \subseteq B$, then $Q(A, B')$

In this part of the introduction, I summarized the basic formal notions of quantifiers. Formal studies on quantifiers have produced several influential ideas that inspired the logical perspective in cognitive studies. In the next section, I will summarize the most important assumptions of this view.

1.1.2 The logical view on quantifiers

The Generalized Quantifier Theory constitutes a starting point and main theoretical framework for the experimental semantics studies on representations and verification models of quantifiers. I will refer to the perspective taken by these studies as the *logical perspective* on quantifiers. The experimental semantics studies that took the logical perspective made three crucial assumptions. Firstly, they adopted the truth-conditional definitions of quantifiers from the Generalized Quantifier Theory. Secondly, they postulated to investigate the verification of quantifiers as a step-by-step procedure of computing the truth value of the sentence based on truth-conditional representation. Thirdly, they tried to link the formal properties of quantifiers with the constrains on the human cognition. In this section, I will give examples of how these assumptions have influenced the studies on quantifiers and point out some shortcomings of logical perspective.

In the logical perspective view, quantifiers are specified via truth conditions (cf. van Tiel et al., 2021). Some quantifiers are context-independent (e.g., *none*, *all*), while the meanings of others vary across contexts (e.g., *many*, *few*). In EXAMPLE 1.1.2 (3), I provided the truth-conditional definition of *many*. The definition refers to the threshold *d*, but does not specify it. Barwise and Cooper (1981) proposed the *fixed context assumption* according to which all determiners, including *many*, can be precisely interpreted in a rich fixed context[6]. While different theoretical proposals (e.g., Barwise & Cooper, 1981; Fernando & Kamp, 1996) have been put forward to incorporate context dependency into quantifier

[6]There are alternative approaches to the Barwise and Cooper's (1981) proposal. For example, Fernando and Kamp (1996) proposed the analysis of *many* in terms of expectations.

theory, the individual differences in quantifier representations have gained little attention. Few existing studies have shown that language users differ in their representations of context-dependent quantifiers (e.g., Yildirim, Degen, Tanenhaus, & Jaeger, 2016).

Moreover, it is not only context-dependent quantifiers that can have various representations, but also quantifiers traditionally treated as context independent, such as *most*. Following the Generalized Quantifier Theory, the first experimental semantics studies on *most* (e.g., Hackl, 2009; Pietroski et al., 2009; Lidz, Pietroski, Halberda, & Hunter, 2011) assumed that it is truth-conditionally equivalent to *more than half*. However, the evidence from other studies (Kotek, Sudo, & Hackl, 2015; Denić & Szymanik, 2020) showed that *most* may have different truth conditions than *more than half*.

To investigate cognitive representations of quantifiers, I propose to refrain from the strong assumption about fixed truth-conditional interpretations of quantifiers. Moreover, I suggest enriching the logical perspective on quantifiers by assuming that individuals can have different representations of logical words. The individual differences seem crucial for understanding variation in mental representations of quantifiers. In this thesis, I will provide a proposal on how to investigate the individual differences in experimental semantics.

The second influential idea in the logical perspective was to directly link the representation of quantifiers to verification procedures. In this view (e.g., Hackl, 2009; Lidz et al., 2011), the truth condition of a quantifier specifies the *algorithm* to assess the truth value of the sentence. The verification process can be split into procedural steps to compute the meaning of the quantified sentence.

The idea that the truth-conditional representation of quantifiers guides the verification process influenced the vast majority of experimental semantic studies (e.g., Hackl, 2009; Pietroski et al., 2009; Lidz et al., 2011). Nonetheless, this idea struggles to capture inconsistency in the experimental data. The first example comes from the paradigmatic case of *most* and *more than half*. Both quantifiers can have multiple equivalent truth-conditional definitions. In EXAMPLE 1.1.2 (2), I specified the truth conditions for *most*. However, the same formula can be also applied to *more than half*. Let us reconsider the EXAMPLE 1.1.4 from the previous subsection. In this example, I specified the truth condition of the sentence "*Most* dogs bark." According to the logical view, the same truth condition holds for the sentence "*More than half* of the dogs bark." Moreover, we can formulate an alternative, yet truth-conditionally equivalent, logical form of this sentence (cf. Hackl, 2009):

1.1.9. Example. $|[\![DOG]\!] \cap [\![BARK]\!]| > 1/2|[\![DOG]\!]|$

While according to the Generalized Quantifier Theory both formulations in EXAMPLES 1.1.4 and 1.1.9 are equivalent, they are not cognitively equivalent. According to the logical perspective (cf. Hackl, 2009; Pietroski et al., 2009, 2011),

each logical form *triggers* a verification strategy. To verify the quantifier based on logical form in EXAMPLE 1.1.4, one has to compute the number of dogs that bark and the number of dogs that do not bark, and compare them. The sentence is true if the number of dogs that bark is greater. The logical form in EXAMPLE 1.1.9 triggers a different verification strategy than in EXAMPLE 1.1.4. To verify the sentence, one has to compute the number of dogs that bark and half of all dogs, and then compare them.

Together, the logical perspective assumes that the truth-conditional representation of quantifiers guides the cognitive process of quantifier verification. However, this view does not provide a comprehensive explanation of how individuals choose one procedure over another. While some studies argued that some strategies are preferred for specific quantifiers (Pietroski et al., 2009, 2011), others showed that people use multiple strategies (Talmina et al., 2017). The ability of individuals to adopt multiple strategies and adjust the verification procedure to the context of the task can not be explained on the basis of the logical view. In this thesis, I will investigate how the verification process depends on the type of task (sentence vs. picture verification), the numerical information provided in the task (proportion), and the participant's individual representation of the quantifier (threshold).

Verification is influenced not only by the truth conditions, but also by other formal properties of quantifiers, such as monotonicity. Barwise and Cooper (1981) formulated a famous prediction that the verification of the downward monotone quantifiers should take longer than that of upward monotone quantifiers due to a difference in the verification procedures. Let us consider an intuitive example. To verify the sentence "*More than 20* students passed the exam," one has to find a so-called *witness set*, namely at least 21 students who passed, to judge the sentence as true. Once the witness set is established the verification process is finished. In contrast, to verify a sentence with the downward monotone quantifier "*Fewer than 20* students passed the exam," one has to check the whole set of students to determine whether they passed. Under the assumption that the set of students is greater than 21, the verification of *fewer than 20* should take more time.

Barwise and Cooper's (1981) prediction inspired a number of experimental studies on the monotonicity effect (e.g., Szymanik & Zajenkowski, 2013; Deschamps et al., 2015; Agmon et al., 2019). The inconsistency in predictions about the monotonicity effect arose for proportional quantifiers. Szymanik and Zajenkowski (2013) argued that participants use the witness set strategy only for numerical quantifiers (*more than seven* or *fewer than eight*), but not for proportional quantifiers (e.g., *more than half* or *fewer than half*). In contrast, Grodzinsky, Agmon, Snir, Deschamps, and Loewenstein (2018) predicted that participants will also use the witness set strategy for proportional quantifiers. The difference in predictions could be due to different assumptions about the truth-conditional representations of proportional quantifiers.

Again, let us consider the EXAMPLES 1.1.4 and 1.1.9. Let us formulate the

truth condition for the sentence "*Fewer than half of the dogs bark.*":

1.1.10. Example. $|[\![DOGS]\!] \cap [\![BARK]\!]| < |[\![DOGS]\!] - [\![BARK]\!]$

1.1.11. Example. $|[\![DOGS]\!] \cap [\![BARK]\!]| < 1/2|[\![DOGS]\!]|$

The representation in EXAMPLE 1.1.10 is not compatible with the witness set strategy, as it requires a comparison of the proportion of dogs that bark to dogs that do not. Therefore, all dogs have to be checked for both *more than half* and *fewer than half*. For the representation in EXAMPLE 1.1.11, the witness strategy can be applied. This is because for *more than half* the strategy requires comparing the witness set of dogs that bark to half of all dogs and, therefore, the verification is finished as soon as the witness set is sufficiently large. For *fewer than half*, in turn, establishing the numeracy of the witness set is not sufficient, but it is also necessary to check all dogs to determine if they bark. As in the case of *most* and *more than half*, also in the monotonicity example, the logical perspective does not appear to capture the flexibility in verification strategies.

The last assumption of the logical perspective is that quantifiers share common properties across natural languages. The mental representations of quantifiers are constrained by the truth conditions as well as formal properties of quantifiers. Base on the Generalized Quantifier Theory, it has been proposed (Barwise & Cooper, 1981) that formal properties of quantifiers, such as conservativity, isomorphism, or monotonicity, constitute *semantic universals*. Experimental semantics studies aimed to establish how the formal, linguistic constraints are related to cognitive constraints. While the formal properties of quantifiers are well-defined on the basis of the Generalized Quantifier Theory, the linking assumption is not provided. A number of studies tried to test different linking assumptions such as the learnability hypothesis (Steinert-Threlkeld & Szymanik, 2019; Hunter & Lidz, 2013), the complexity of quantifiers (van de Pol, Steinert-Threlkeld, & Szymanik, 2019; van de Pol, Lodder, Maanen, Steinert-Threlkeld, & Szymanik, 2021), and communication pressure (e.g., cultural evolution, Carcassi, Steinert-Threlkeld, & Szymanik, 2019).

To summarize, the logical perspective on quantifiers has several limitations. Firstly, it does not account for individual differences in meaning representations. In this thesis, I will argue that individual differences in linguistic behavior are widespread (cf. Kidd, Donnelly, & Christiansen, 2018) and unavoidable in the realm of quantifiers. Secondly, the link between the mental representations of quantifiers and verification is often unequivocal. The logical perspective does not predict which verification strategy should be used in a specific context. It does not account for the flexibility in verification strategies. Finally, the logical perspective assumes a link between formal constraints on the natural language quantifiers and constraints on humans cognitive abilities. The nature of this link is, however, still a matter of debate. To build a cognitively realistic model of

quantifiers processing, the shortcomings of logical perspective have to be overcome. The studies should incorporate the individual differences into quantifiers representations, allow modelling flexibility in quantifier verification in different contexts, and establish empirically the link between formal properties of quantifiers and human cognitive abilities. In the next section, I will outline a new, *cognitive perspective* on quantifiers that attempts to fill the gaps of logical view.

1.2 Toward cognitive perspective on quantifiers

In the logical perspective view, the starting point for studies on quantifiers is the unquestioned quantifier representations in a form of truth condition derived from the Generalized Quantifier Theory. The goal of empirical studies is to test participants' behavior during verification under the assumption that they have access to the truth-conditional representations predicted by linguistic analysis and that there are no individual differences in behavior.

In this thesis, I propose a different approach, namely the *cognitive perspective* on quantifiers. The cognitive view takes as its starting points both the truth-conditional quantifier representations derived from the Generalized Quantifier Theory and participants' behavior during verification. In this view, the truth-conditional representations of quantifiers are treated as hypotheses about cognitive representations. These hypotheses are directly tested against the experimental data in the verification task. The behavior in the verification task is described by the *cognitive model*, which has to fulfil several requirements. The model cannot include fixed truth-conditional representations, but it should be able to recover quantifier representations from the data. Moreover, it should allow the representations to differ between quantifiers and individual participants. It can include several parameters to account for different aspects of quantifier representations such as truth conditions assigned to quantifiers by participants (*thresholds*) or vagueness. In this way, the cognitive perspective accounts for individual differences in truth-conditional representations and verification procedures.

In this thesis, I will use different computational models to link the experimental data with the representations of quantifiers. The computational models are the implementation of the cognitive model of the quantifier verification task. They can be used to test the predictions of the cognitive model. They should allow mapping between the cognitive processes and different aspects of semantic representations. For example, the model could distinguish the semantic representation of a quantifier from participants' certainty about the representation. They should also account for different formal properties of quantifiers such as monotonicity. Moreover, the computational models ought to account for individual differences in semantic representations and task performance (e.g., mistakes during the verification task). Finally, they should make it possible to investi-

gate the step-by-step verification process of quantifiers by extracting the stages of processing.

The need to study linguistic behavior by means of computational models has already been recognized (cf. Frank & Goodman, 2012; Degen & Tanenhaus, 2019; van Tiel et al., 2021; Schlotterbeck, 2017). In the domain of quantifiers, we can distinguish two main modeling approaches: semantic automata and Bayesian models. In the following sections, I will argue that these two modeling traditions are not well-suited to the goal of the cognitive model and I will propose an alternative approach.

1.2.1 Current computational models of quantifiers

The semantic automata models describe verification of quantifiers on a very abstract level. For each quantifier, they define a formal language and an abstract computing device (machine) that decides on a given input (a representation model/ situation) if the quantifier is true or false (van Benthem, 1986). To pursue computation, the machine goes through several states and changes the state according to the instructions. The complexity of the automata model depends on the quantifier complexity. In the semantic automata framework, the processing of quantifiers is modelled as step-by-step computation. Experimental studies have linked the reaction times associated with quantifier verification to the complexity of computations performed by the machine (Szymanik & Zajenkowski, 2013, 2010) and to verification strategies (Steinert-Threlkeld, Munneke, & Szymanik, 2015). Moreover, McMillan, Clark, Moore, Devita, and Grossman (2005) linked the complexity of computing devices to cognitive processes such as working memory (see also Szymanik & Zajenkowski, 2010; Zajenkowski et al., 2014). Together, the semantic automata models were successful in explaining the mean differences in reaction times during verification of quantifiers of varying complexity (Szymanik & Zajenkowski, 2010). However, their application is limited. For example, the semantic automata model can not convincingly explain the monotonicity effect[7].

The growing popularity of the Bayesian models in experimental semantics and pragmatics research is evidenced by the increasing number of studies using these models to account for various linguistic phenomena, such as behavior in a truth value judgment task (e.g., Waldon & Degen, 2020), choice of quantifier as a scene description (e.g., van Tiel et al., 2021), and differences in the use of truth-conditionally equivalent quantifiers (e.g., Carcassi & Szymanik, 2021). The Bayesian modeling tradition extends beyond linguistic models. The main assumption of Bayesian models is that behavior is based on predictions. Language users compute meanings by applying the Bayes' Rule and inferring the message of other language users. While the Bayesian inference is successful in explaining the

[7]Specifically, the automata models can account for the monotonicity and truth value interaction, but not for the main effect of monotonicity.

linguistic behavior patterns, the mechanism of meaning computation is described on an abstract level and does not refer to the cognitive processes (such as attention, working memory, and executive functions) *per se*. The Bayesian models are very successful in modeling *what* the language users represent, but not *how* they represent it (Martin, 2016).

The semantic automata and Bayesian models do not satisfy all requirements of the cognitive model of quantifiers. The advantage of semantic automata models is that they model verification as a process and map stages of the verification process onto different machine states. The disadvantage is that they neither model the individual differences in verification nor account for flexibility in meaning representations (Steinert-Threlkeld et al., 2015). The Bayesian models, in turn, are successful in explaining the individuals' choices in a given context, but not in explaining the cognitive processes behind them. In this thesis, I will leave these two modeling traditions aside and use computational models taken from experimental psychology which can account for all the requirements of the cognitive model.

1.2.2 New modeling approach

Because the existing models of quantifiers are not entirely suitable for the purpose of this thesis, I sought a different modeling tradition. In this thesis, I will use various computational models such as the Diffusion Decision Model (Ratcliff, 1978), an evidence accumulation model of decision-making from mathematical psychology, the hidden semi-Markov model multivariate pattern analysis model (Anderson, Zhang, Borst, & Walsh, 2016), and a three-parameter logistic regression model inspired by Item Response Theory to build the cognitive model of quantifier verification[8].

To provide an impression of the new modeling approach, I will briefly discuss how these models can account for quantifier verification based on one example. I will use the Diffusion Decision Model (Ratcliff, 1978) to model the quantifier representations and the monotonicity effect. The Diffusion Decision Model satisfies the requirements of the cognitive model of quantifiers. It is a processing model that accounts for reaction times and response data. It also allows for individual differences by estimating parameters for each participant. Moreover, it can map model parameters onto different cognitive processes (semantic representation or performance), and vary parameters across quantifiers to model formal properties.

Furthermore, the Diffusion Decision Model is an example of the evidence accumulation model. It originated from the memory retrieval model proposed by Ratcliff (1978). It assumes that the decision process in the memory retrieval task is the process of comparing two representations: the representation stored in

[8]The detailed motivation of modeling choices is provided together with models' descriptions in the relevant chapters.

memory and the stimuli representation (probe). This process is described as the accumulation of the evidence in favor of the hypotheses that the two representations match or mismatch. The speed of this process depends on the distribution of matching and mismatching features. In mathematical terms, evidence accumulation is a continuous random walk process called the diffusion process. The evidence accumulation process in the Diffusion Decision Model can be linked to the quantifier verification task. The representations of quantifiers are stored in participants' memory and compared to the experimental input in the verification task. The process of comparison is analogous to memory retrieval. In conclusion, we can link the model describing a general process of evidence accumulation to a specific task such as quantifier verification and link the cognitive process of decision-making described by the model to the verification process. By setting the model parameters, we can further study how this process depends on the properties of different quantifiers.

Computational models from the nonlinguistic domain have been gaining increasing attention as models of linguistic phenomena (Martin, 2016). This trend has two advantages. Firstly, the computational models provide a unified framework and allow modeling language-specific and domain-general processes together. The second advantage is that, by using the computational models taken from experimental psychology, we also advance the modeling field. For example, the Diffusion Decision Model (Ratcliff, 1978) was very successful in explaining behavior in simple two-choice decision tasks such as: recognition memory tasks (e.g., Ratcliff, 1978), random dot-motion task (e.g., Palmer, Huk, & Shadlen, 2005), speed-accuracy trade-off task (e.g., van Maanen, Portoles, & Borst, 2021; Katsimpokis, Hawkins, & van Maanen, 2020), lexical decision task (e.g., Tillman, Osth, van Ravenzwaaij, & Heathcote, 2017; Ratcliff, Thapar, & McKoon, 2010), and non-symbolic number comparison task (e.g., Ratcliff et al., 2010; Park & Starns, 2015; Kang & Ratcliff, 2020) as well as many others (see Ratcliff & McKoon, 2008; Ratcliff, Smith, Brown, & McKoon, 2016, for review). The simple model of the two-choice task can be extended to capture a wider variety of phenomena. Successful attempts can be found in the numerical cognition field (Ratcliff & McKoon, 2018). I believe that both the linguistics and modeling fields can benefit from incorporating concepts from theoretical linguistic into popular models from experimental psychology.

1.3 Thesis overview

The goal of this thesis is to provide a cognitive model of quantifiers. This model should account for the formal properties of quantifiers and individual differences between language users. To develop such a model, I studied quantifiers in various verification tasks and explored the topics previously investigated from the logical perspective. In the next section, I will give a brief overview of the thesis content.

1.3.1 How are quantifiers ordered?

In Chapter 2, I explored the individual differences in quantifier representation by using the three-parameter logistic model. Moreover, I used the machine learning method of clustering to establish groups of participants with different mental representations of quantifiers. I established three groups of participants with a different meaning representations of *most*, *many*, and *few*. In addition, I investigated the order of quantifiers on the mental line. Numbers are strictly ordered on a number line. Quantifiers also seem to be ordered according to the information they convey about magnitudes which (Pezzelle et al., 2018). I showed that three clusters of participants organize quantifiers differently on the mental line.

1.3.2 How are quantifiers represented and processed?

In Chapter 3, I tested the difference in representations of *most* and *more than half*. I applied the computational model to account for different aspects of these representations, such as truth conditions and vagueness. Moreover, I provided a framework to test the individual differences in quantifier representations. I also tested how the choice of representation affects the quantifier verification process. I found substantially greater variability in meaning representations of *most* than *more than half*. Moreover, I found that the verification of *most*, but not *more than half*, is proposition dependent. The findings challenged the logical theories' assumption of truth-conditional equivalence of *most* and *more than half*.

1.3.3 Why are some quantifiers more difficult to process than others?

Linguistic analysis (Barwise & Cooper, 1981) predicts that the downward entailing quantifiers should be processed slower than upward entailing quantifiers. This prediction has been borne out in many experimental studies (e.g., Grodzinsky et al., 2018; Deschamps et al., 2015; Agmon et al., 2019; Just & Carpenter, 1971). I refer to this finding as a polarity effect[9]. Reaction times are the most widespread measure of the polarity effect. In the previous experiments, the mean reaction times from different experimental conditions were compared to indicate the difficulties of processing negative quantifiers (e.g., *fewer than half*) compared to positive quantifiers (e.g., *more than half*). Nonetheless, the measure of the mean reaction times has a strong limitation. It only provides information about the average cognitive process duration. By using the trial-by-trial information, we

[9]The negative polarity notion refers to the idea that a negative expression in natural language contains a hidden negation. Negative polarity could apply to quantifiers as well as to other natural language expressions such as adjectives. The notion of monotonicity refers to entailment patterns and is characteristic of quantifiers, but not adjectives (see Agmon et al., 2019, for discussion). In this thesis, these notions will sometimes be used interchangeably.

can zoom into the stages of processing of the quantified sentences and establish the source of the differences in mean reaction times .

In Chapters 4 and 5, I addressed different explanations of the polarity effect. I tested two competing accounts, two-step and pragmatic models, by using computational models and behavioral and electroencephalography data. Furthermore, I zoomed into the processing stages of the quantified sentence. While the behavioral data presented in Chapter 4 support both two-step and pragmatic accounts, the electroencephalography data analysis presented in Chapter 5 challenged the two-step model.

1.3.4 How do formal properties of quantifiers link to human cognitive abilities?

It has been postulated that natural languages share some common properties, called universals (Barwise & Cooper, 1981). The challenge for experimental semantics is to establish how the formal properties of quantifiers link to human cognitive abilities. I refer to this challenge as finding a *linking assumption*. It has been proposed that universal properties facilitate learning (Steinert-Threlkeld & Szymanik, 2019). This is referred to as the *learnability hypothesis*. The learnability hypothesis was tested in the domain of quantification in many modeling studies (e.g. Steinert-Threlkeld & Szymanik, 2019; Carcassi, Steinert-Threlkeld, & Szymanik, 2019), as well as in a few experimental studies (Lidz et al., 2011; Spenader & de Villiers, 2019). In Chapter 6, I addressed the learnability hypothesis from the experimental perspective. I showed that learnability can explain only some of the semantic universals. I also discussed the number of methodological challenges related to the large-scale experimental investigation of semantic universals.

Chapter 2

Most quantifiers have *many* meanings[1]

Abstract Logical theories of meaning assume that function words, such as natural language quantifiers, have a fixed meaning expressed by their truth conditions. In this study, we challenge this view by showing that there are systematic individual differences in semantic representations of quantifiers. Using computational modeling, we separated three sources of individual differences: truth condition, vagueness, and response error, and mapped them on different model parameters. We selected five natural language quantifiers (*few*, *fewer than half*, *many*, *more than half*, and *most*), which we expected to differ in the model parameters. We collected response data in an online experiment and fitted a Bayesian three-parameter logistic regression model. By applying the k-means clustering algorithm to the model's parameters, we found three subgroups of participants with different semantic representations of quantifiers and the organization of the mental line of quantifiers. Moreover, we found asymmetry between positive and negative quantifiers in response error and vagueness. This finding supports the view that logical words, like content words, are sensitive to individual differences, and hence it challenges the logical theories of meaning.

2.1 Introduction

Needless to say, humans differ in their cognitive abilities. Similar to other cognitive domains, individual differences are also present in natural language processing (Kidd et al., 2018). In this paper, we investigate individual differences in natural language quantifier representations. Natural language quantifiers make an excellent case study as they have drawn the attention of researchers from different fields ranging from logic (Barwise & Cooper, 1981; Mostowski, 1957) to formal semantics (Keenan & Paperno, 2012; Szabolcsi, 2010) to cognitive science

[1]This chapter is based on the manuscript: Ramotowska, Haaf, van Maanen, and Szymanik (2022), *Most* quantifiers have *many* meanings (unpublished manuscript).

(Ramotowska et al., 2020b see Szymanik, 2016 for review). Quantifiers, such as *many*, *few*, *most*, *some*, and *at least 5*, are used to express quantities. They belong to the close class of functional words. They have been studied mostly in the verification paradigm (e.g., Deschamps et al., 2015; Hackl, 2009; Pietroski et al., 2009; Schlotterbeck, Ramotowska, van Maanen, & Szymanik, 2020), in which participants have to decide if a sentence containing quantifiers is true in a given context.

Individual differences have been studied extensively in many natural language domains, including gradable adjectives (Verheyen, Dewil, & Égré, 2018) and semantic categorizations of nouns (Verheyen & Storms, 2013; Verheyen et al., 2018; Verheyen, White, & Égré, 2019). While the studies of quantifiers are common, individual differences in their use have gained somewhat less attention in the literature. The dominant perspective on quantity words considers meaning representations to be logical forms (Hackl, 2009; Pietroski et al., 2011). A growing body of evidence (Denić & Szymanik, 2020; Ramotowska et al., 2020b; Talmina et al., 2017) questions this traditional view and calls for incorporating individual differences in the domain of quantifiers.

Individual differences in quantifiers may come from three different sources. The first source are differences in general cognitive abilities, e.g., working memory (Just & Carpenter, 1992; Kidd et al., 2018) or executive functions (Kidd et al., 2018). For example, the accuracy and speed of verification of proportional quantifiers depend on working memory capacities (Steinert-Threlkeld et al., 2015; Zajenkowski & Szymanik, 2013; Zajenkowski et al., 2014) and cognitive control (Zajenkowski & Szymanik, 2013; Zajenkowski et al., 2014). The second source of individual differences could lay in the choice of verification strategies (Talmina et al., 2017). For example, Talmina et al. (2017) showed that some participants prefer to use a precise strategy while verifying *most*, and others choose an approximate strategy. Moreover, strategy preference depends on the context (Register, Mollica, & Piantadosi, 2018). Finally, the third source of individual differences could be different semantic representations of quantifiers where individuals assign different truth values to the same sentence (Spychalska, Kontinen, & Werning, 2016). Spychalska et al. (2016) divided participants into two groups based on their truth value evaluation of the underinformative sentence "*Some* As are B" when in fact *all* As were B. The group of so-called pragmatic responders judged the underinformative sentence as false and logical responders as true.

While the first two sources of individual differences are compatible with the formal semantics perspective on language, the last one contradicts the intuition that language users have to agree on the truth condition of the sentences in order to communicate. At first glance, it seems that rational subjects cannot assign different meanings to logical words such as quantifiers. Nevertheless, in this paper, we show that the last option is tangible. We aim to answer three questions regarding individual differences in quantifiers. First, *how many subgroups of participants with different meanings can we identify*? Second, *how are the meanings*

of quantifiers interrelated at the subject level? Third, we want to separate behavioral and semantic sources of individual variation in quantifier representations. We considered the truth conditions (quantifier's threshold) and vagueness as the semantic source. We include a response error parameter representing mistakes that participants made during the verification process to account for behavioral sources. Response errors could happen due to attentional lapses or difficulties in processing of complex quantifiers but are unrelated to vagueness or thresholds. The third question regarding individual differences is therefore: *How are the parameters in our model interrelated*?

To answer these questions, we analyzed data from a quantifier verification task, in which participants were asked to judge the truth of a quantified sentence based on information about proportion. We modeled the choices using a logistic regression model and estimated three model parameters corresponding to threshold, vagueness, and response errors. Then, we clustered participants based on the parameter estimates. Computational modelling has previously been successfully applied to test competing semantic theories (van Tiel et al., 2021) and to distinguish between different sources of individual differences in language processing (Vasishth, Nicenboim, Engelmann, & Burchert, 2019; Waldon & Degen, 2020). Moreover, computational modelling allows the investigation of qualitatively different effects in experimental data (Haaf & Rouder, 2019; Donzallaz, Haaf, & Stevenson, 2021; Kolvoort, Davis, van Maanen, & Rehder, 2021; Miletić & van Maanen, 2019; Ramotowska, Steinert-Threlkeld, van Maanen, & Szymanik, 2020a). Our work continues the tradition of using computational modeling to better understand cognitive representations. In the following section, we explain the reasons for each of our questions and modeling choices.

2.1.1 How many subgroups of participants with different meanings can we identify?

The logical theory of meaning (e.g., Generalized Quantifier Theory, Barwise & Cooper, 1981; Mostowski, 1957) analyses the meaning of quantifiers in terms of truth conditions. The natural language quantifier's truth condition specifies a threshold above or below which the quantifier is true[2]. For example, the quantifier *most* in the sentence "*Most* of the As are B" is true ($most(A, B) = 1$), if the intersection of sets A and B ($|A \cap B|$) is greater than the intersection of sets A and not B ($|A \cap \neg B|$). Example 2.1.1 shows truth conditions for quantifiers: *most*, *more than half*, *fewer than half*, *many*, and *few*.

[2]In this paper, we focus only on quantifiers with one threshold. Some quantifiers can have two or more thresholds, e.g., *between 3 and 6* has two thresholds, 3 and 6.

2.1.1. Example.

1. *Most* $(A, B) = 1$ iff $|A \cap B| > |A \cap \neg B|$
2. *More than half* $(A, B) = 1$ iff $|A \cap B| > |A|/2$
3. *Fewer than half* $(A, B) = 1$ iff $|A \cap B| < |A|/2$
4. *Many* $(A, B) = 1$ iff $|A \cap B| > n$, where n is some cardinality or proportion
5. *Few* $(A, B) = 1$ iff $|A \cap B| > n$, where n is some cardinality or proportion

Some quantifiers like *at least 5* have clear truth conditions with the threshold equals 5. Other quantifiers, like *many*, have various thresholds depending on the context (Schöller & Franke, 2016). Moreover, *many* and *few* are ambiguous between cardinal and proportional reading (Partee, 1989). According to cardinal reading, the threshold is a fixed number e.g., "*Many* students passed the exam" means more than 40 students. Proportional reading of *many*, in turn, refers to *many* as more than some proportion, e.g., "*Many* of the students passed the exam" means more than 40% of the students. In this paper, we focus only on proportional readings of *few* and *many*.

Individual differences seem likely in context-dependent quantifiers such as *many* and *few*. Yildirim et al. (2016) showed that different speakers have different meanings of these quantifiers. More surprisingly, Ramotowska et al. (2020b) found individual differences in the quantifier *most* within the experimental paradigm downplaying the role of context. This finding questions the underlying assumption of many studies (Hackl, 2009; Pietroski et al., 2009; Lidz et al., 2011) that participants have a dominant representation of *most*. In the current paper, we performed a cluster analysis to systematically investigate the subgroups of participants.

2.1.2 How are the meanings of quantifiers interrelated at the subject level?

The meanings of the quantifiers considered here highly overlap. They constitute the sets of alternatives for each other. The first studies that looked into the order of quantifiers on a scale tried to link quantifiers with proportions for psychometric purposes (Hammerton, 1976; Newstead, Pollard, & Riezebos, 1987). They found that participants were less consistent in the usage of some quantifiers than others. For example, low-magnitude quantifiers were more context-dependent than high-magnitude quantifiers (Newstead et al., 1987).

Recently, Pezzelle et al. (2018) have shown that quantifiers can be ordered on the mental number line. However, the distance between meaning representations

does not have to be equal (see also van Tiel et al., 2021). For example, low-magnitude quantifiers (e.g., *few*, *almost none*) were more separated from each other and had sharper representations than high-magnitude quantifiers (*almost all*, *most*, *many*). They also showed that some quantifiers are semantically more similar than others. For example, *many* is more similar to *most* than to *few*. Moreover, the change in the meaning representation of one quantifier (e.g., *many*) affects the threshold of the polar opposite quantifier (e.g., *few*, Heim et al., 2015). This effect is present in the reinforcement learning paradigm (Heim et al., 2015) or via adaptation during exposure (Heim et al., 2020).

The above studies did not account for the individual differences in quantifier meaning representation. In contrast, we investigated the relationship between quantifier meanings taking into account the between-subjects variability in thresholds to shed more light on how quantifiers are represented on the mental number line on the individual level.

2.1.3 How are the parameters of our model interrelated?

Vagueness

Quantifiers such as *many* and *few* are vague, which means that their meaning boundaries depend on the situation (Newstead & Coventry, 2000; Solt, 2011). Another characteristic of vagueness concerns the borderline cases. If we agree that the sentence "*Many* of the students failed the exam." is true when 20% of students failed, we will also probably agree that the sentence is true when 19% failed. Thus, the threshold for accepting a statement as true for *many* and *few* is fuzzy even given a fixed context (Solt, 2011).

Some studies showed that the quantifier *most* is also vague (Denić & Szymanik, 2020; Solt, 2011). Solt (2016) claimed that *most* and *more than half* are represented on different underlying scales. *More than half* has to be represented on the ratio scale, while *most* requires only the semiordered scale. The latter scale allows less precise comparisons, and, therefore, the meaning of *most* is more variable. Moreover, Denić and Szymanik (2020) showed that participants were less consistent about their threshold for *most* than for *more than half*. Taken together, context dependency is not the only factor that might change the quantifier threshold. In a fixed context, some quantifiers can have variable truth condition assignments due to vagueness. Therefore, we included a separate parameter in our model to test the effect of vagueness independently of the threshold.

Response error

While verifying quantified sentences, participants sometimes make errors. The response error in quantifier verification tasks depends on quantifier complexity (Zajenkowski & Szymanik, 2013), working memory demands (Zajenkowski &

Szymanik, 2013), or polarity (Zajenkowski & Szymanik, 2013; Deschamps et al., 2015). For example, participants process negative quantifiers slower and with a higher error rate than when they process positive quantifiers (Just & Carpenter, 1971; Deschamps et al., 2015; Schlotterbeck et al., 2020).

Moreover, previous studies (Hackl, 2009) argued that the same overall proportion of errors in the verification task for *most* and *more than half* speaks in favor of the same truth conditions of these quantifiers. In contrast, another study (Kotek et al., 2015) showed that the accuracy for *most* is lower than for *more than half* when the proportion is slightly above 50%. Kotek et al. (2015) interpreted this asymmetry as a difference in quantifier pragmatics rather than truth conditions. Finally, Denić and Szymanik (2020) showed that the accuracy for *most* is lower than for *more than half* relative to their estimated thresholds. These studies show that the response error is a crucial measure of participants' performance. However, its interpretation is not unequivocal. We included the additional response error parameter in our model to account for differences in accuracy between negative and positive quantifiers and to disentangle the measure of error from the measures of threshold and vagueness.

To summarize, even though the above discussion suggests that vagueness, threshold, and error may be interrelated, as far as we know, this relationship has not been systematically investigated on an individual level. For example, we can imagine that participants may have the same truth conditions for *most* and *more than half* and yet perform worse while verifying *most* because of other reasons. Moreover, participants may make more errors when verifying vague quantifiers. Response errors and vagueness, in turn, can lead to variability in thresholds. These interdependencies might lead to confounds when interpreting the experimental data. Therefore, we applied a model with three different parameters to capture these three aspects.

2.1.4 Current study

To test the individual differences in quantifier representations and the relationship between the meanings of different quantifiers, we asked participants to judge the truth of a sentence involving a quantifier against the proportion given as a number between 1% and 99%. We chose proportional quantifiers from three groups: quantifiers with sharp meaning boundaries (*fewer than half* and *more than half*); vague and context-dependent quantifiers (*few* and *many*); and one quantifier that falls between these groups (*most*). After fitting a computational model to the response data to estimate these parameters for every quantifier and participant, we performed a cluster analysis on the threshold parameter to establish the subgroups of participants with different meanings. We predicted that all participants would have the same threshold for *fewer than half* and *more than half* because these quantifiers already refer to the threshold, namely half. In contrast, we predicted that we would find between-clusters variability in thresholds

for vague quantifiers like *most*, *many*, and *few*. We also hypothesized that only vague quantifiers would contribute to clustering on the threshold.

Moreover, to address our second research question, we explored how the meaning of one quantifier relates to other quantifiers. Firstly, we tested the correlations between thresholds on the group level to see if the thresholds between quantifiers are interrelated. In contrast to previous studies (Hammerton, 1976; Heim et al., 2015; Newstead et al., 1987; Pezzelle et al., 2018; van Tiel et al., 2021), we also looked into the order of quantifiers on a mental scale on the individual level within the clusters of participants.

Finally, we tested the relationship between model parameters. We wanted to separate the between-participants variability in truth conditions (thresholds) from vagueness and response error by introducing three parameters into our model. We tested whether the model parameters were correlated. We did not have specific predictions about the direction of these correlations. This analysis was exploratory in nature. Nonetheless, we predicted a higher value of the vagueness parameter for vague quantifiers and that participants would make more mistakes while verifying the negative quantifiers. In addition to clustering on threshold parameters, we performed a cluster analysis on vagueness and response error to see which quantifiers contributed to clustering. We expected that *few*, *many*, and *most* would contribute to clustering on vagueness and negative quantifiers to clustering on response error.

Before running the computational model, we explored the effects of the three parameters on potential data patterns. In particular, we wanted to separate vagueness and response error effects because they both lead to response variability. Response errors are a result of additional cognitive processes and should therefore occur after the participants compare the proportion given in the experimental trial to their internal threshold. As such, response errors are independent of proportion. In contrast, vagueness adds noise to the decision process. The noise is greater around the participants' threshold. As a result, the internal threshold shifts from trial to trial. As such, vagueness depends on the proportion.

Figure 2.1 presents how we conceptualized threshold, response error, and vagueness parameters. We chose the quantifier *more than half* for illustration. For the ideal responder, the proportion of 'true' responses below 50% is zero, and above 50% is one. The logistic curve has a sharp shape indicating a rapid shift from false to true responses at the threshold. When the responses are affected by vagueness, the perceived threshold varies from trial to trial, and the logistic curve increases gradually. The response error, in turn, does not change the shape of the response curve. Instead, it lowers the probability of the true response above the threshold and increases the probability of the true response below the threshold equally for all proportions. We also plotted the combined effect of response errors, vagueness, and threshold.

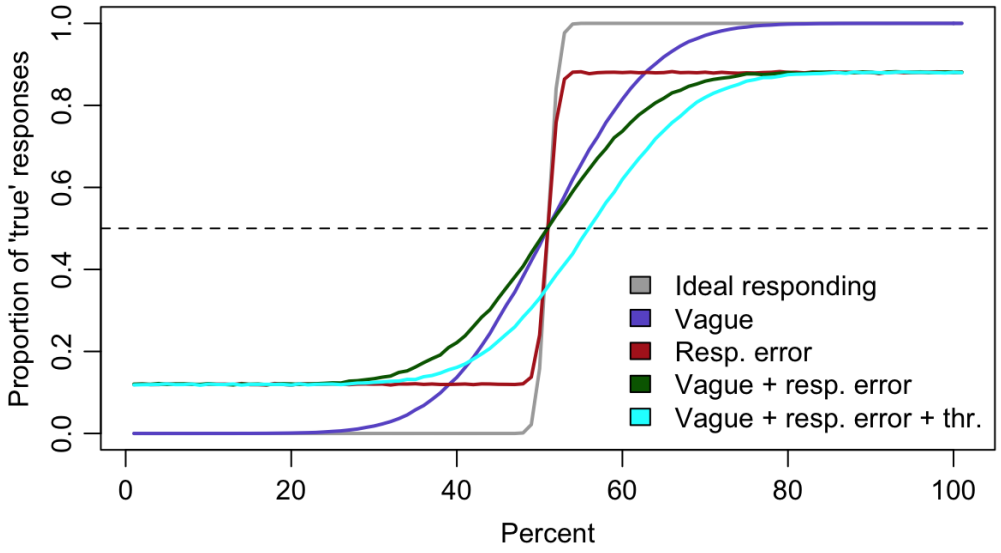


Figure 2.1: Predicted logistic curves under different threshold (thr.), response error (resp. error), and vagueness (vague) parameters. The dashed line indicates the 0.5 proportion of true responses. The percentage for which the logistic curve crosses the dashed line is the threshold.

2.2 Method

2.2.1 Participants

We recruited 90 participants via the online recruitment platform Amazon Mechanical Turk. We excluded 19 participants based on three exclusion criteria. Firstly, we excluded 11 participants who had 50% or more reaction times faster than 300 ms. Secondly, we excluded 7 participants who failed to obey the monotonicity of quantifiers. We defined the monotonicity criterion in the following way: for positive quantifiers (*many*, *most*, and *more than half*) we expected the probability of providing the true response to increase with increasing proportion. The opposite effect should hold for negative quantifiers. To apply monotonicity criterion, we fitted the generalized linear model to participants' response data with the proportion as a predictor and with by-subject random intercept and slope for proportion (*glmer* R function, Kuznetsova, Brockhoff, & Christensen, 2017). We excluded participants, who had a negative slope for positive quantifiers or a positive slope for negative quantifiers. Finally, we excluded 1 participant, who took part in a similar experiment. These exclusions meant that we included 71 participants (47 male, age $M = 35$, range: 22–59) in the final sample.

2.2.2 Experimental Design and Procedure

In our experiment, participants had to indicate whether the sentence with the quantifier: *most*, *many*, *few*, *fewer than half*, or *more than half* was true or false based on the sentence containing a proportion ranging from 1% to 99% (excluding 50%). We did not include the proportion 100%, because Ariel (2003) showed that *most* has an upper bound on meaning and using it with 100% proportion is not accepted, although it is highly accepted with 99%. The upper bound of most could cause a divergence in the logistic function which we used in our model. We did not include 50%, because this proportion could be confusing for *more than half* and *fewer than half*.

While *most*, *more than half* and *fewer than half* have a proportional interpretation (Hackl, 2009), as explained above, *many* and *few* are ambiguous between cardinal and proportional reading (Partee, 1989). For example, *many* could mean more than a certain number (cardinal reading) or more than a certain proportion (proportional reading, see Example 2.1.1). We used explicit partitive 'of the' and present proportions as a percentage for all quantifiers to ensure the proportional reading and avoid confounds for ambiguous quantifiers. Moreover, by using the percentage format we enforced the precise comparison between proportion and the threshold. In this way, we minimized the differences between quantifiers in verification strategies. For example, in some experimental paradigms *most* is verified using approximate strategy (Pietroski et al., 2009), while in others mixtures of strategies is used (Talmina et al., 2017).

The experiment started with a short training block to familiarize participants with the procedure. Next participants completed the 250 trials (50 per quantifier) in randomized order. At the end of the experiment, participants provided basic demographic information. Each trial of the experiment consisted of two sentences displayed on separate screens. The first sentence containing the quantifier was of the form "[*Most/Many/Few/More than half/Fewer than half*] of the gleerbs are fizzda." To read this sentence participants had to press the arrow down key and keep it pressed. When they advanced to the next screen, they read a sentence containing proportion e.g., "20% of the gleerbs are fizzda." Participants had to provide a response by pressing the right or left arrow keys corresponding to true or false judgment (counterbalanced between participants).

In our experiment, we used pseudowords generated from 50 English six-letters nouns and adjectives using *Wuggy* software (Keuleers & Brysbaert, 2010). We used pseudowords to avoid pragmatic effects associated with quantifiers. The original words were controlled for frequency (*Zipf* value 4.06, van Heuven, Mandera, Keuleers, & Brysbaert, 2014). A native English speaker assessed the pseudowords in terms of how well they imitated English words.

2.2.3 Data pre-processing

We excluded trials with response times shorter than 300ms and longer than 2500ms (similar cut-offs to Ratcliff & McKoon, 2018). Altogether, we excluded 6% of trials. To be able to fit the same logit model to all quantifiers we flipped the true and false responses for *few* and *fewer than half*.

2.2.4 Computational Model

The logistic regression model is suitable for modelling the threshold variability (Ramotowska et al., 2020b). The model assumes that the probability that participants verify a statement as true or false depends on the proportion that was presented on a particular trial and the values of the logistic function parameters asymptote, midpoint and scale:

$$response \sim \frac{asymptote}{1 + exp(midpoint - proportion)/scale} \tag{2.1}$$

To accommodate individual differences and differences between quantifiers in the model, we used a three-parameter logistic regression model inspired by Item Response Theory (IRT). IRT determines the relationship between an individual's trait and the probability of providing a correct response for a given item (Hanlbleton, Swaminathan, & Rogers, 1991; Ligia et al., 2013). This relationship is expressed by the Item Response Function, which maps the IRT parameters (difficulty, discrimination, and guessing) onto the logistic function. The three-parameter model has a difficulty parameter, which determines the level of an individual trait necessary to provide a correct response (midpoint), a discrimination parameter that determines the steepness of the logistic curve (scale), and a guessing parameter that can adjust the logistic curve asymptotes.

In our model, the threshold corresponds to the difficulty parameter, vagueness to the discrimination parameter, and response error to the guessing parameter from the IRT model. We used a hierarchical Bayesian model to estimate the parameters for each participant-quantifier combination. To fit the model, we used the *rstan* package in R (Stan Development Team, 2017) with 6 chains, 750 warm up iterations per chain and 2500 iterations per chain.

The model was specified in the following way. Let i indicate participants, $i = 1, ..., I$, j indicate the quantifier, $j = 1, ..., 5$, and k indicate the trial for each quantifier, $k = 1, ..., K_{ij}$. Then Y_{ij} is the i-th participant's response to the j-th quantifier in the k-th trial, and $Y_{ijk} = 1$ if participant indicated true, and $Y_{ijk} = 0$ if participant indicated false. Then, we may model Y_{ijk} as a Bernoulli, using the logit link function on the probabilities:

$$Y_{ijk} \sim Bernoulli(\pi_{ijk}) \tag{2.2}$$

where the probability space of π maps onto the μ.

$$\pi_{ijk} = \gamma_{ij} + (1 - 2\gamma_{ij}) logit^{-}1(\mu_{ijk}) \tag{2.3}$$

The additional parameter γ_{ij} determines the probability of making a response error on either side of the threshold, namely erroneously saying true, or erroneously saying false. Each participant-quantifier combination has its own response error parameter estimate. The parameter μ_{ijk} has a linear model explication:

$$\mu_{ijk} = \frac{c_{ijk} - \beta_{ij}}{\alpha_{ij}} \tag{2.4}$$

where c_{ijk} indicates the percentage centered at 50%, parameters β_{ij} indicate the threshold, and parameters α_{ij} correspond to the vagueness of the quantifier.

We defined prior probabilities on response error (γ), threshold (β), and vagueness (α) parameters:

$$\gamma_{ij} \sim Beta(2, 20) \tag{2.5a}$$

$$\beta_{ij} \sim Normal(\delta_j, \sigma_j^2) \tag{2.5b}$$

$$\alpha_{ij} \sim log - Normal(\nu_j, \sigma_{\alpha_j}^2) \tag{2.5c}$$

$$\nu_j \sim Normal(0, 5^2) \tag{2.5d}$$

$$\sigma_{\alpha_j}^2 \sim Invers - Gamma(2, 0.2) \tag{2.5e}$$

$$\sigma_j^2 \sim Invers - Gamma(2, 0.2) \tag{2.5f}$$

$$\delta_j \sim Normal(0, 5^2) \tag{2.5g}$$

The hierarchical nature of the distributions for α_{ij} and β_{ij} indicate that we estimated the effect of threshold and vagueness for each participant under the assumption that they had a common mean and variance. The vagueness and threshold priors were fairly uninformative. Vagueness (α_{ij}) came from a log-normal distribution to ensure only the positive estimates. Its mean (ν_j) had a normal distribution, and its variance ($\sigma_{\alpha_j}^2$) was drawn from Inverse-Gamma distribution, as this distribution is typically used to model variance. For the thresholds (β_{ij}) we used a normal distribution with a common, normally-distributed mean (δ_j) and the same variance distribution (σ_j^2) as for α_{ij}. The response error (γ_{ij}) came from a more informed distribution with most of its mass below an error rate of 20% for each true and false response[3].

[3]To reduce the complexity of the model, we did not use hierarchical modeling for response errors.

2.2.5 Cluster analysis

We ran the exploratory cluster analysis for threshold, vagueness and response errors separately, estimating the clusters using the K-means clustering method (*kmeans* function in R, Hartigan & Wong, 1979). We determined the optimum number of clusters by using the elbow plots and Silhouette width.

2.2.6 Linear Discriminant Analysis

To assess the contribution of the model estimates to the clustering, we performed a linear discriminant analysis (LDA). We used the stepwise procedure Wilks' lambda assessment (*greedy.wilks* function in R package *klaR*, Roever et al., 2015) to determine which variable contributed significantly to cluster formation. Next, we ran the LDA (*lda* function in R package *MASS*) to test how accurately the selected variables could predict the clusters. To validate the LDA, we ran a leave-one-out cross validation.

2.3 Results

2.3.1 Estimated parameters

The estimated model parameters are shown in Table 2.1. Figure 2.2 shows the estimated item response curves for each participant-quantifier combination; the overall response curves for the quantifiers are represented by the bold, colored lines. We found greater individual variation in thresholds for *most*, *many* and *few*, compared to *more than half* and *fewer than half*. At the group level, quantifier thresholds were represented in the following order (Friedman test $\chi^2(4) = 134$, $p < 0.001$, moderate effect size $W = 0.47$): *few* had the lowest threshold, followed by *many*, then were *fewer than half* and *more than half*, and *most* had the highest threshold (pairwise comparison, Wilcoxon Signed Rank Test with Bonferroni correction).

The quantifiers *fewer than half* and *more than half* were the least vague as indicated by the steep response curves in Figure 2.2. Moreover, *few* was more vague than *fewer than half* ($V = 2556$; $p < 0.001$), *many* was more vague than *more than half* ($V = 2556$; $p < 0.001$), *many* was more vague than *most* ($V = 2556$; $p < 0.001$), and *most* was more vague than *more than half* ($V = 2556$; $p < 0.001$), p - values based on Wilcoxon Signed Rank Test. We also found that *fewer than half* had a greater response error than *more than half* ($V = 2323$; $p < 0.001$), and *few* had greater response error than *many* ($V = 1809$; $p = 0.002$), p - values based on Wilcoxon Signed Rank Test. As predicted, the vague quantifiers had a higher value of vagueness parameter and negative quantifiers had higher value of response error parameter.

Table 2.1: Mean (*SD*) parameters for each quantifier, and additionally for threshold parameter the percent corresponding to mean thresholds.

	Threshold	Vagueness	Response error
Few	-.103 (.073), 39.7%	.016 (.001)	.062 (.042)
Fewer than half	-.006 (.027), 49.4%	.002 (.00004)	.074 (.047)
Many	-.061 (.094) 43.9%	.019 (.003)	.048 (.024)
More than half	.001 (.012) 50.1%	.001 (.00003)	.042 (.019)
Most	.029 (.056) 52.9%	.009 (.001)	.047 (.024)

In the next step, we studied the associations between model parameters across quantifiers to reveal potentially systematic patterns (see Figure 2.3). Figure 2.3a shows the correlations between thresholds. These correlations were negligible or weak. This finding gives reason for the cluster analysis, because the lack of correlation might be caused by different relationship between thresholds in the subgroups. It also suggests that clusters of participants could have different representation and ordering on the mental line.

Figure 2.3b shows the correlations for vagueness, and Figure 2.3c for response errors. The correlations for vagueness were also weak, suggesting that this parameter is quantifier-specific and not domain-general. In contrast, the correlations for response error varied, ranging from a strong correlation between *few* and *fewer than half* ($r = 0.75$), to the weakest correlation between *more than half* and *many* ($r = 0.24$, see Figure 3C). The strongest correlation was significantly higher than the weakest, Stringer's test $z = 4.72$, $p < 0.001$. This suggests that response error reflects general cognitive ability.

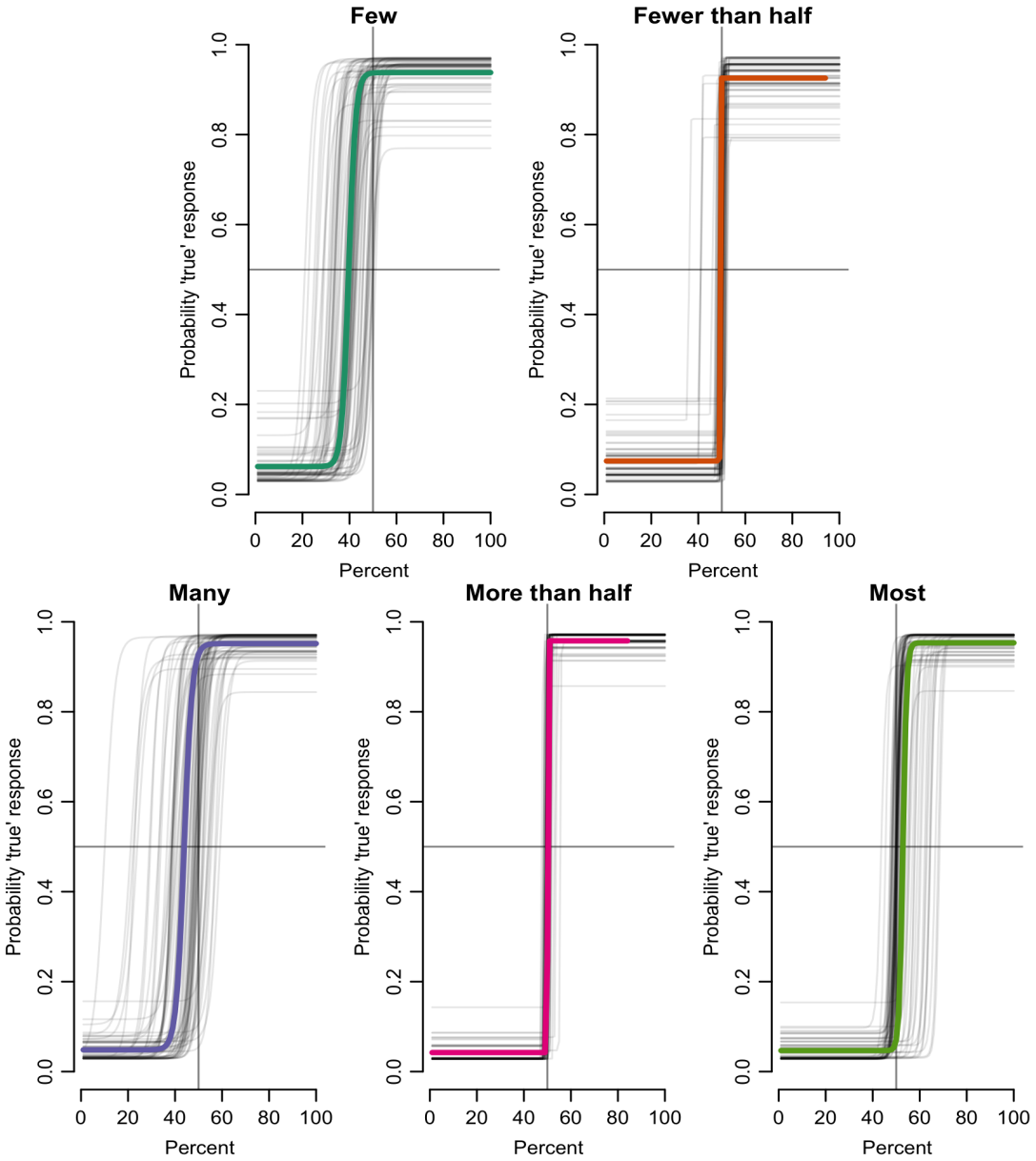


Figure 2.2: The logit curves estimated for each quantifier. The color lines indicate the mean curves.

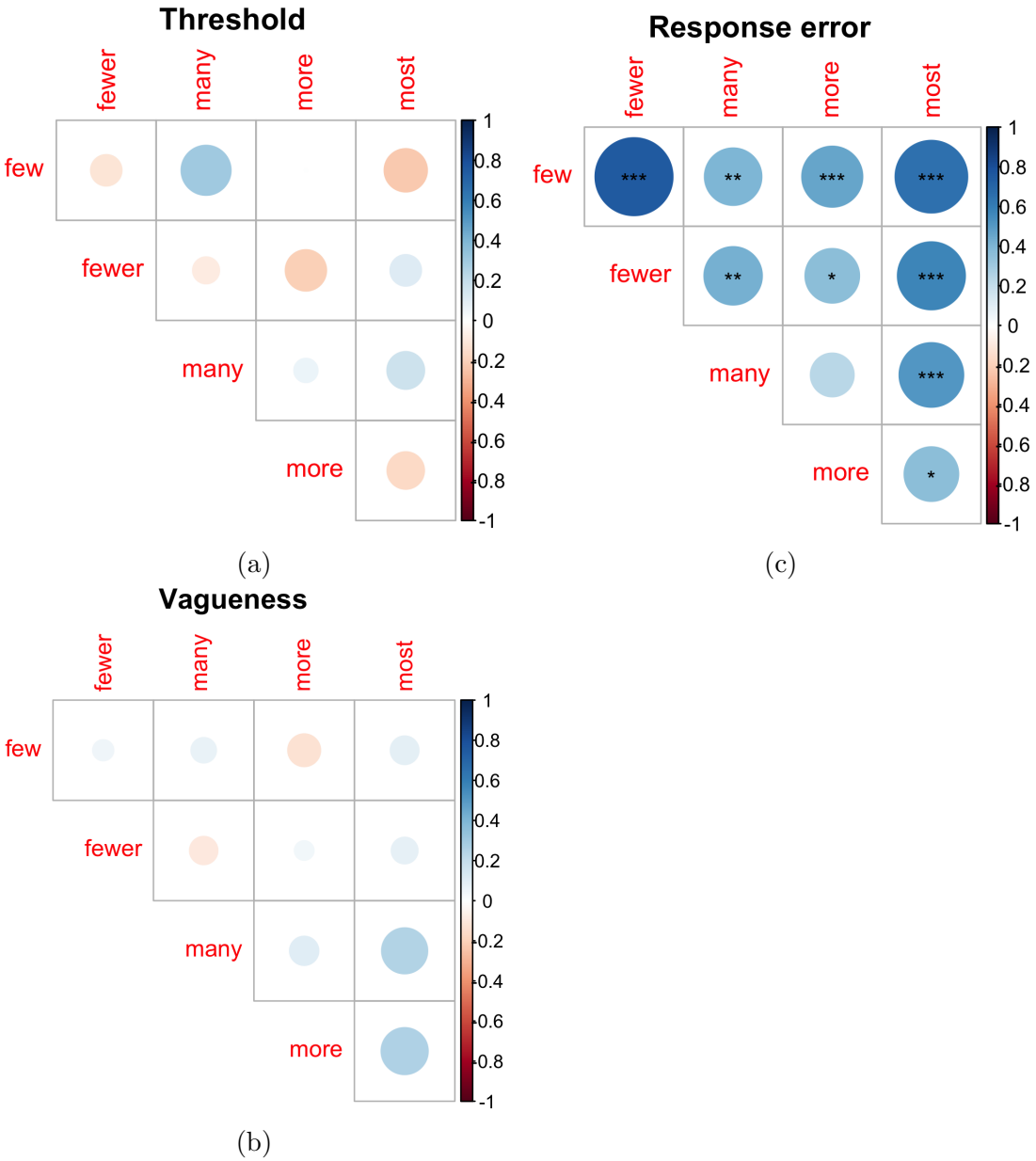


Figure 2.3: 2.3a Correlations between thresholds; 2.3b correlations between vagueness; 2.3c correlations between response error (significance level *** 0.001, ** 0.01,* 0.05). The p values were adjusted using the Bonferroni correction.

To test the interrelationship between vagueness, threshold, and response error, we correlated the model parameters for each quantifier (Figure 2.4). This correlation analysis was exploratory in nature. We wanted to test whether there were any systematic patterns across quantifiers. We found a significant negative correlation between threshold and vagueness for *few* (r = -0.33) and *many* (r = -0.31). We also found correlations between threshold and response error for *fewer*

than half (r = -0.32), and response error and vagueness for *many* (r = 0.53) and *most* (r = 0.52). In general, the correlations did not reveal systematic patterns. The lack of systematic correlations between vagueness and response error parameters gives additional support to the choice to model these parameters as two separate mechanisms.

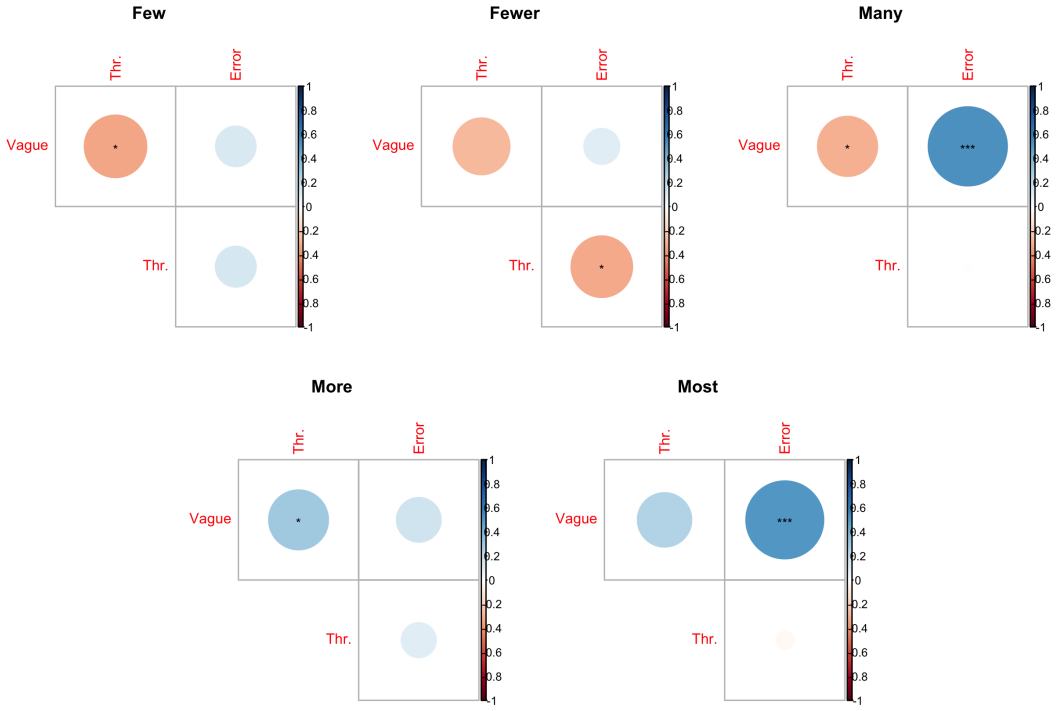


Figure 2.4: Correlations of parameters for each quantifier (significance level *** 0.001, ** 0.01, * 0.05). The p values were adjusted using the Bonferroni correction.

2.3.2 Cluster analysis results

Threshold

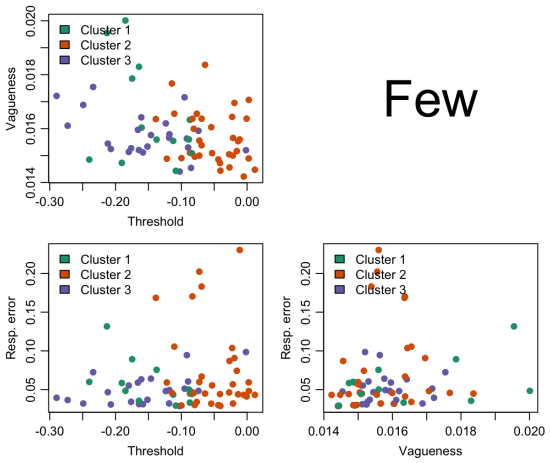
The methods to determine the optimum number of clusters for threshold gave ambiguous results. The elbow plot indicated 3 or 4 clusters, while the Silhouette method preferred 5 clusters. We chose the simplest solution, comprising 3 clusters, because the additional clusters consisted of only 4 participants, making interpretation difficult. The three clusters were indistinguishable for the quantifiers *fewer than half* and *more than half*, but differed substantially in thresholds for the quantifiers *few*, *many*, and *most*. Figure 5 shows the individual estimates for threshold, vagueness, and response error parameters for the quantifiers *few*, *many*, and *most*, with color indicating cluster membership.

The first cluster ($N = 13$) consisted of participants with a higher mean threshold for *most*, the second cluster ($N = 34$) included participants who had thresholds for all quantifiers close to 50%, and the last cluster ($N = 24$) consisted of participants who had similar a mean threshold for *few* and *many* (see Table 2.2). In addition, we found that participants in Cluster 3 had a higher tendency to make errors, with this tendency especially visible for few (see Figure 5).

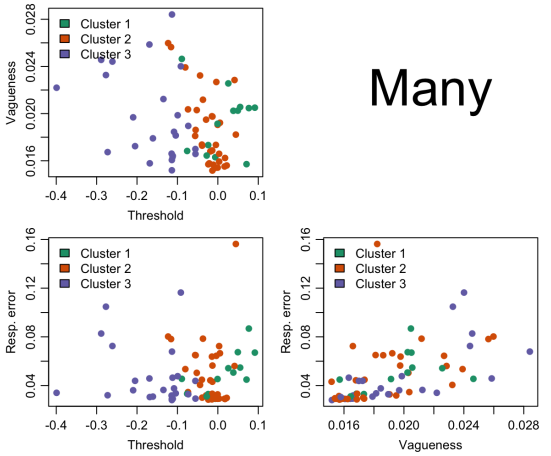
Table 2.2: Mean (SD) threshold parameter in each cluster and percentage corresponding to mean thresholds, 3-cluster solution.

Quantifier	Cluster 1 ($N = 13$)	Cluster 2 ($N = 34$)	Cluster 3 ($N = 24$)
Few	-.15 (.05)	-.05 (.04)	-.15 (.07)
	35%	45%	35%
Fewer than half	.001 (.01)	-.012 (.03)	-.002 (.02)
	50.1%	48.8%	49.8%
Many	.014 (.06)	-.022 (.04)	-.16 (.09)
	51.4%	47.8%	34%
More than half	-.00006 (.006)	.002 (.01)	.0007 (.01)
	49.99%	50.2%	50.07%
Most	.10 (.05)	.009 (.03)	.02 (.05)
	60%	50.9%	52%

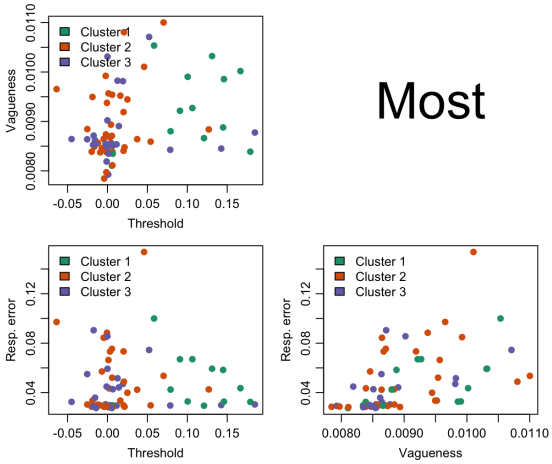
Because we did not find a systematic relationship between thresholds of different quantifiers (see Figure 2.3a), we investigated this relationship in the clusters (see Figure 2.6). We supposed that the lack of correlations between thresholds could be explained by the different relationships between quantifiers in subgroups. Specifically, we wanted to test whether all participants would have the same order of vague quantifiers on a mental line and whether the distance between quantifiers would be different in clusters. Figure 2.6a shows that all participants had a lower or equal thresholds for *many* than for *most*. However, the distance between thresholds was higher in Cluster 3 than in other clusters. Figure 2.6b shows that the vast majority had a higher threshold for *many* than for *few*. The greatest distance between thresholds was in Cluster 1, while the smallest was in Cluster 3. Figures 2.6c and 2.6d show that all participants in Cluster 3 had a lower threshold for *many* than for *more than half* and *fewer than half*.



(a)



(b)



(c)

Figure 2.5: Relationship between threshold, vagueness, and response error for *few* (2.5a), *many* (2.5b), and *most* (2.5c), indicating three clusters based on threshold. Cluster 1 ($N = 13$) is indicated in green, Cluster 2 ($N = 34$) in orange, Cluster 3 ($N = 24$) in purple.

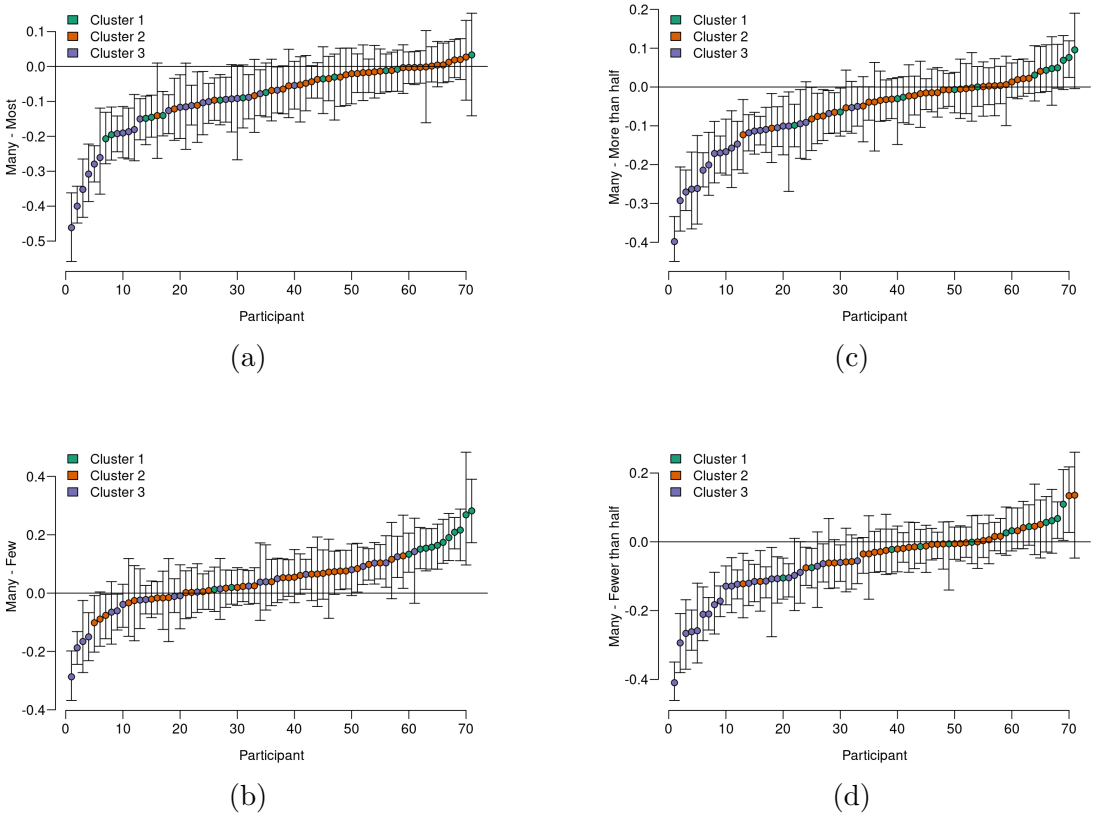


Figure 2.6: 2.6a The difference between the threshold for *many* and *most* for each participant. 2.6b The difference between the threshold for *many* and *few* for each participant. 2.6c The difference between the threshold for *many* and *more than half* for each participant. 2.6d The difference between the threshold for *many* and *fewer than half* for each participant. Colors are used to indicate cluster membership: Cluster 1 is indicated in green ($N = 13$), Cluster 2 in orangne ($N = 34$), and Cluster in purple ($N = 24$). The error bars indicate the 95% credible intervals.

Vagueness

The elbow plot and Silhouette method agreed that the two-cluster solution was optimal, identifying one cluster ($N = 24$) with high vagueness for *many*, and a second cluster ($N = 47$) with lower vagueness for *many* (Table 2.3). We expected polar opposite quantifiers *few* and *many* to make comparable contributions to clustering on vagueness. What we observed instead was the asymmetry in *many* and *few*. Figure 2.7 shows that participants with higher vagueness for *many* had a tendency to make more mistakes and had lower threshold, while participants with lower vagueness for *many* had a threshold concentrated around 50% and made fewer errors.

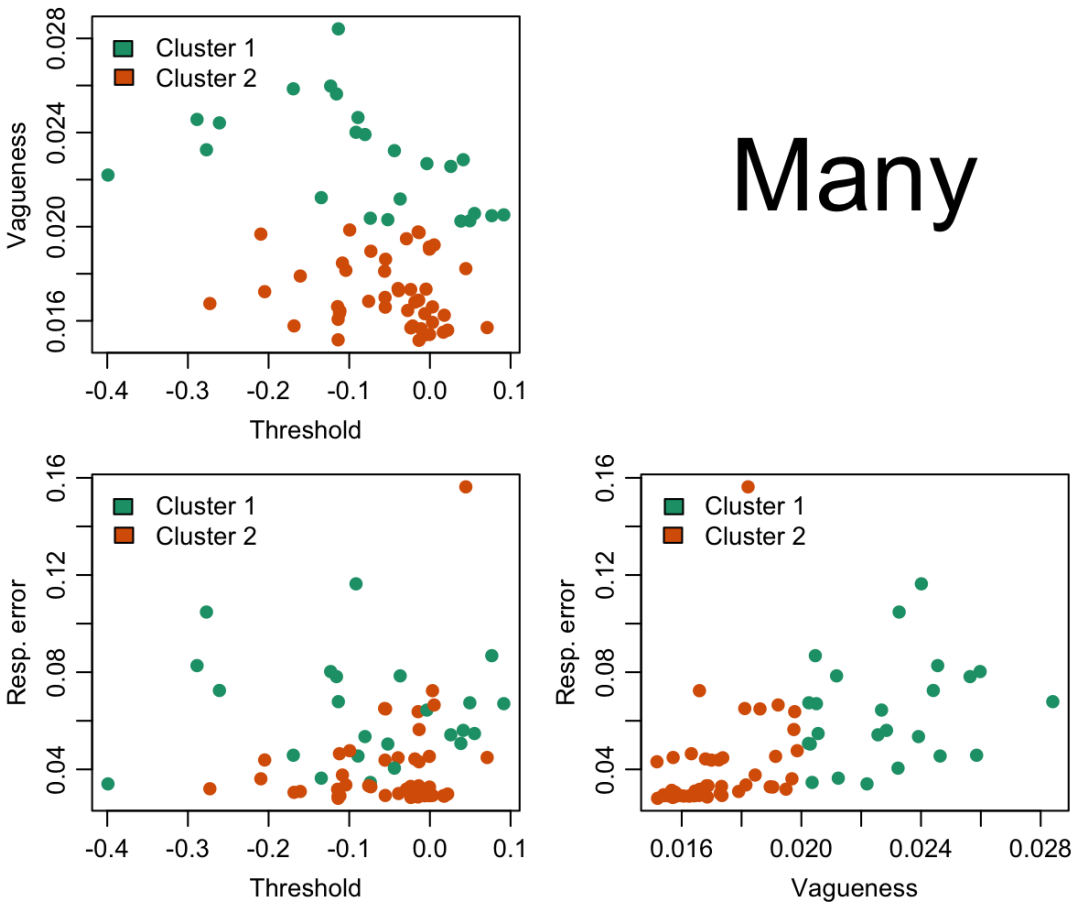


Figure 2.7: Relationship between threshold, vagueness, and response error for *many*, indicating two clusters based on vagueness. Cluster 1 ($N = 24$) with higher vagueness for *many* is indicated in green, and Cluster 2 ($N = 47$) with lower vagueness for *many* in orange.

Table 2.3: Mean (SD) vagueness parameter in each cluster, 2-cluster solution.

Quantifier	Cluster 1 ($N = 24$)	Cluster 2 ($N = 47$)
Few	.016 (.001)	.016 (.001)
Fewer than half	.002 (.00004)	.002 (.00004)
Many	.023 (.002)	.017 (.001)
More than half	.001 (.00004)	.001 (.00002)
Most	.009 (.001)	.009 (.001)

Response error

The elbow plot suggested that either two or three clusters should be optimal, but the Silhouette method indicated the 2-cluster solution. Assuming two clusters, we found a cluster of participants with few response errors ($N = 64$) and a cluster with more response errors ($N = 7$) across quantifiers, see Table 2.4. This means that the majority of participants had a low response error rate. The difference in response error between clusters was most prominent for negative quantifiers. Figure 2.8 shows the relationship between model parameters based on response error clustering for *few* and *fewer than half*. For *few*, we did not observe that participants with a high response error had a tendency toward more extreme thresholds or vagueness, while for *fewer than half* some participants that made more errors also had lower threshold.

Table 2.4: Mean (SD) response error parameter in each cluster, 2-cluster solution.

Quantifier	Cluster 1 ($N = 7$)	Cluster 2 ($N = 64$)
Few	.17 (.05)	.05 (.02)
Fewer than half	.19 (.03)	.06 (.03)
Many	.08 (.04)	.05 (.02)
More than half	.06 (.02)	.04 (.02)
Most	.09 (.03)	.04 (.02)

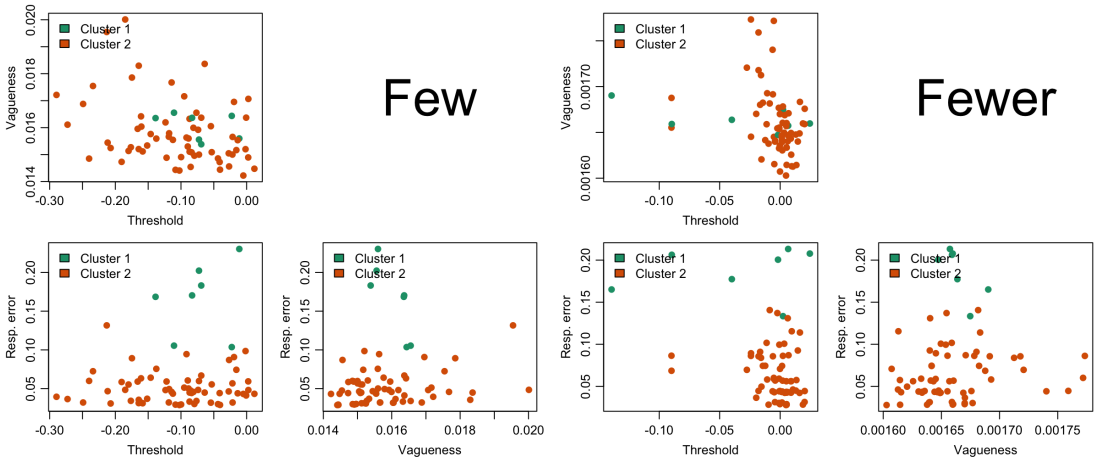


Figure 2.8: Relationship between threshold, vagueness, and response error for *few* (2.8a) and *fewer than half* (2.8b), indicating two clusters based on response error. Cluster 1 ($N = 7$) with a high response error is indicated in green, and Cluster 2 ($N = 64$) with a lower response error in orange.

2.3.3 Linear Discriminant Analysis results

Threshold

For thresholds, as expected, we found that only vague quantifiers contributed to the clustering: *many* ($\lambda = 0.42$, $p < 0.001$), *few* ($\lambda = 0.24$, $p < 0.001$), and *most* ($\lambda = 0.16$, $p < 0.001$). Figure 2.9 shows the combined effect of the three quantifiers on the clustering. The LDA accuracy in classification into Clusters 1 to 3 based on thresholds for *many*, *few* and *most* was 97%, and the leave-one-out cross validation accuracy was 94%.

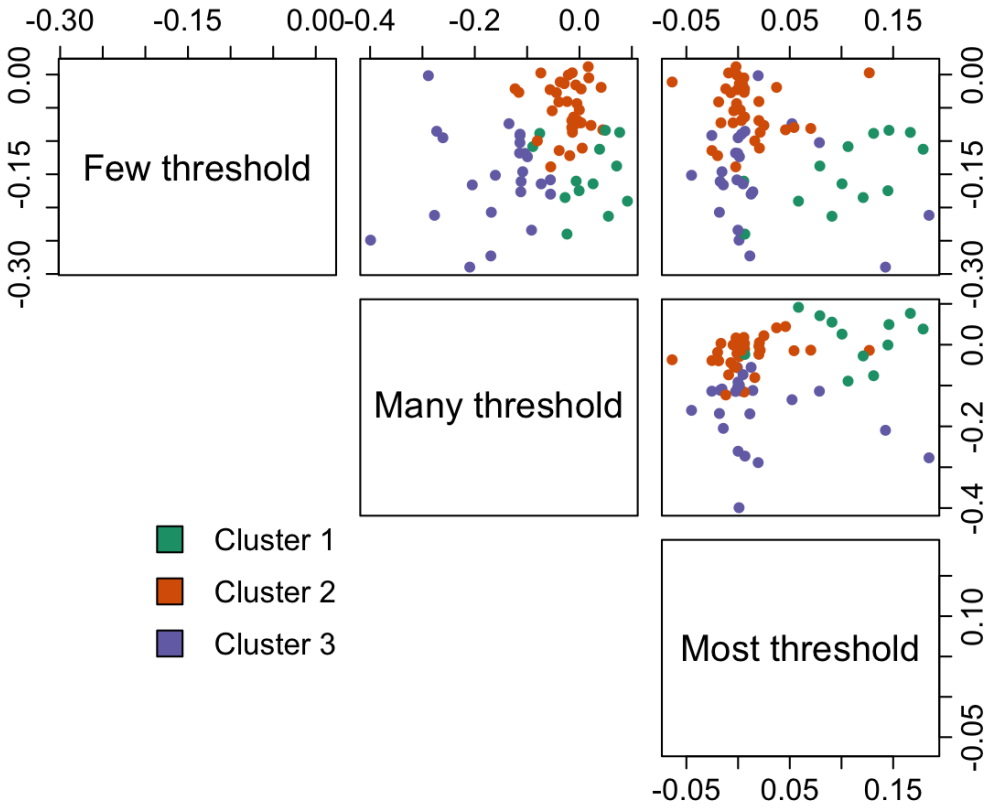


Figure 2.9: Three clusters for threshold based on *few*, *many*, and *most* parameters. The parameters' values of thresholds for three quantifiers (*few*, *many*, and *most*) that contributed to clustering are plotted against each other. Colors are used to indicate the cluster membership: Cluster 1 ($N = 13$) is indicated in green, Cluster 2 ($N = 34$) in orange, and Cluster 3 ($N = 24$) in purple.

Vagueness

For the vagueness parameter, we expected vague quantifiers to contribute to the clustering. We found that only *many* contributed significantly to the clustering ($\lambda = 0.29$, $p < 0.001$). The LDA achieved 94% accuracy in classification of participants into clusters based on vagueness parameters for *many*, and the leave-one-out cross validation accuracy was 94%.

Response error

We expected the response error parameter for negative quantifiers to contribute more to clustering. In line with this hypothesis, the Wilks test showed a significant contribution of response error parameters for *few* ($\lambda = 0.32$, $p < 0.001$) and *fewer than half* ($\lambda = 0.25$, $p < 0.001$), but not for *many*, *most* and *more than half*. Figure 2.10 shows the combined effect of the two quantifiers on clustering. Participants who made more errors while verifying *few* also made more errors for *fewer than half*. We used the LDA to predict the cluster membership for each participant based on response error parameters for *few* and *fewer than half*. The LDA achieved 99% accuracy, and the leave-one-out cross validation accuracy was 99%.

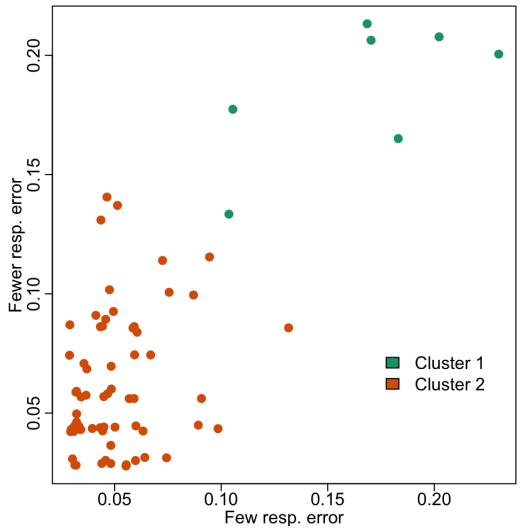


Figure 2.10: Two clusters for response error based on *few* and *fewer than half* parameters. The response error values of parameters for *fewer than half* are plotted against the response error values for *few*. Colors are used to indicate the cluster membership: Cluster 1 ($N = 7$) with high response error is in indicated in green, and Cluster 2 ($N = 64$) with low response error in orange.

2.4 Discussion

Previous studies showed that quantifiers are organized on a mental scale (Hammerton, 1976; Pezzelle et al., 2018) and that participants use their internal threshold to verify proportional quantifiers (Shikhare et al., 2015). However, little has been known about the individual differences in the organization of quantifiers on the mental line. The main goal of this study was to identify the subgroups of participants with different meanings of quantifiers. We investigated how quantifiers are organized on the mental line within the subgroups.

Firstly, we examined the correlations between quantifiers for each parameter of our model. We found that only the response errors correlated across quantifiers. The lack of significant correlations for other parameters further motivated the analysis of the subgroups. We ran a cluster analysis on threshold parameters of quantifiers. We identified three groups of participants with different mean thresholds and relationships between the meaning of quantifiers. As initially predicted, quantifiers with sharp meaning boundaries, like *fewer than half* and *more than half*, did not contribute to clustering, and they had similar thresholds in all groups. In contrast, thresholds for *many*, *few*, and *most* varied considerably between clusters. In all groups, *most* had the highest threshold. However, the mean threshold varied between clusters. In the first cluster, the mean threshold was 60%, and in the second and third clusters, the mean thresholds were just slightly above 50%, at 51% and 52%, respectively. For *few*, participants in the first and third clusters had mean thresholds equal to 35%, and in the second cluster, the mean threshold was 45%. The mean threshold for *many* was the most diverse between groups. It ranged from 51% in the first cluster to 48% in the second cluster to and 34% in the third cluster.

The subsequent goal of this paper was to look into the relationship between threshold, vagueness, and response error. As predicted, we found that quantifiers with broad meaning boundaries had a higher vagueness value and that negative quantifiers had a higher response error value. We investigated the correlations across parameters for all quantifiers. However, we failed to find systematic patterns. Finally, we clustered participants based on vagueness and response error. We found two clusters with high and low vagueness for *many* and two clusters with high and low response error for *few* and *fewer than half*. We will discuss the implications of these findings in more detail in the following subsections.

2.4.1 Order of quantifiers on the mental line

Because we failed to find correlations at the group level between thresholds of different quantifiers, we zoomed into the mental line of the subgroups of participants. We observed that the clusters differed in the range of the mental line and the order of quantifiers on it. Participants in the first cluster had the most stretched mental line, ranging between 35% and 60%, with a clear order of thresh-

olds, where *few* was the lowest and *most* was the highest. In contrast, the second group had the most shrunk mental line, ranging between 45% and 51%. The mental scale of the last group stretched between 34% and 52%. We further looked into the relationship between vague quantifier pairs: *few* and *many* (the polar opposites), and *many* and *most*.

Many* vs. *few

Hammerton (1976) found that although participants assigned different numerical equivalence to quantifiers, they were consistent about the order of the quantifiers. Our findings indicate that participants were consistent about the order of some quantifiers, but not all. For example, we found an asymmetry between *many* and *few* with regard to their positioning on the mental scale. The position of *many* on the mental scale was more flexible than the position of *few*. In the second and third clusters, the mean threshold for *many* was lower than for *more than half* and *fewer than half*, but in the first cluster, it was higher (see Figure 2.6). The second asymmetry between *many* and *few* was that only *many* contributed to clustering based on vagueness.

The flexibility of *many* on the mental scale cannot be explained by its context-dependency. Firstly, in our experiment, we used an artificial context by introducing pseudowords. There was no reason for participants to have different expectations about the context.

Secondly, based on the literature (Newstead et al., 1987), we predicted the opposite pattern of results. The low-magnitude quantifiers, such as *few*, are more context-dependent than high-magnitude quantifiers (Newstead et al., 1987). Moreover, they can change their threshold depending on the reference set (Newstead et al., 1987) and they are more separated from each other on the mental scale than high-magnitude quantifiers (Pezzelle et al., 2018).

We attribute the asymmetries in our study to competition between quantifiers. While *few* was less than 50% for all clusters and *most* was more than 50%, *many* had to compete with both quantifiers for a place on the mental line. As a result of this competition, *many* had a greater variation in threshold and was more vague, at least for some participants. We observed two tendencies concerning the threshold of *many* (see Figure 2.6). The first tendency was to either keep the threshold for *few* and *many* close together (Clusters 2 and 3) or far apart (Cluster 1, Figure 2.5b). The second tendency was to either keep the threshold for many close to most (Cluster 2, and to some extent 1) or far from most (Cluster 3, see Figure 2.6a). Despite these tendencies, almost all participants had a higher threshold for *many* than for *few*, and all participants had a higher threshold for *most* than for *many*. Altogether, this finding shows that the position of *many* on the mental line is more flexible than the position of *few* and it explains the membership of the clusters. Nonetheless, in all clusters participants treated *few* as less than *many*, and *many* as less than *most*.

Many vs. *most*

Previous studies and linguistic analysis (Hackl, 2009; Pezzelle et al., 2018) stressed similarities between *most* and *many*. Firstly, Hackl (2009) analyzed *most* as a superlative of *many* (*many*+est). This analysis predicts that *most* has to be more than *many*. Our data support this prediction. We showed that not only the mean threshold for *many* was lower than for *most* in all clusters, but also all participants had a higher threshold for *most* than *many* regardless of the cluster's membership (see Figure 2.6a). While all participants treated *most* as the superlative of *many*, the distance between thresholds of these quantifiers was different depending on the cluster. The greatest distance was in the third subgroup.

Secondly, Pezzelle et al. (2018) showed substantial overlap in the production of *most* and *many*. Both quantifiers cover comparable proportions on the mental scale. In contrast, our results show individual differences in the distance on the mental line between *most* and *many*. For example, in the third cluster, the mean threshold for *many* was considerably lower than the mean threshold for *most*, while in the second cluster, both thresholds were close to 50%.

Lastly, Pezzelle et al. (2018) found that *many* is used less frequently than *most*. We think that the quantifier's vagueness could be one of the sources of the difference in frequency. The high perceived vagueness of *many* lowers its usefulness. The more vague the quantifier, the less information it conveys. However, participants try to be as informative as possible (Grice, 1975) and therefore avoid the usage of uninformative quantifiers with very flexible meanings. This explanation generates a new prediction to test in future work: participants who perceive *many* as vaguer should also use it less often in the production experiment.

2.4.2 Relationship between model parameters

Finally, we tested the relationships between vagueness, thresholds, and response error. We did not find significant correlations for threshold and vagueness between quantifiers, indicating that these parameters were quantifier-specific. In contrast, the response error parameter was significantly correlated across almost all quantifiers. The correlations were, however, stronger between negative than positive quantifiers because of the greater variation in response error in negative quantifiers. Due to this variation, only negative quantifiers contributed to clustering on response error. Response error, thus, reflects a combination of general task performance ability and specific difficulty in verification for negative quantifiers (Deschamps et al., 2015; Just & Carpenter, 1971; Schlotterbeck et al., 2020). We noted that the cluster with a higher rate of response error was small ($N = 7$), probably because the task was generally easy. It would be worth testing whether the response error parameter contributes more to clustering in a more challenging task, for example, with visual displays instead of sentences.

With regard to correlation between parameters for each quantifier, we did

not find systematic patterns for the whole group of participants (see Figure 2.4). The only significant correlation between threshold and response error was for *fewer than half*. This correlation was, however, strongly affected by the outlier participants with a low threshold for *fewer than half* (see Figure A.1 in Appendix A). This finding shows that the variation in thresholds reflects variation in the semantic representations and it is not an artefact of task performance.

The response error correlated positively with the vagueness parameter. However, the correlation was only significant for *many* and *most*. Moreover, participants from the cluster with higher vagueness for *many* also had a higher response error (Figure 2.6). As one could expect, the vaguer the quantifier, the more difficult it is to perform the task. In addition, the lack of systematic correlations between vagueness and response error shows that they correspond to two different processes that should be modeled as separate parameters in the model. The response error reflects the general cognitive mechanism and is affected by a quantifier's difficulty, while vagueness is a semantic property, which may correlate with the verification difficulties of a quantifier (in this study e.g., *most* and *many*), but it can not be equated with a number of errors (cf. Denić & Szymanik, 2020).

Finally, we found significant correlations between vagueness and threshold for *many* and *few*, but, importantly, not for *most*. This finding challenges the explanation proposed by Solt (2016), according to which participants verify *most* using the approximate strategy (Pietroski et al., 2009). Consequently, the verification of *most* is noisy around 50%. To reduce the noise, participants prefer thresholds significantly greater than 50%. This theory predicts that participants with higher thresholds for *most* will perceive it as a vaguer quantifier than subjects with lower thresholds. In our model, we captured the noisiness of verification in the vagueness parameter. The lack of significant correlation between vagueness and threshold for *most* does not support Solt's explanation. Instead, it suggests that some participants assigned different truth conditions to *most* and *more than half*.

2.4.3 Sources of individual differences

Our starting point for considering the individual differences in meaning representations of natural language quantifiers was the observation that language users can have different truth conditions for logical words. For example, previous studies (e.g., Spychalska et al., 2016) showed that two groups of speakers have different interpretations of the quantifier *some*. In this spirit, we demonstrated that this phenomenon is not limited to just one quantifier or to pragmatics. We showed that there are three subgroups of participants with different meaning representations for *many*, *few*, and *most*.

We argue that individual differences are not due to the various verification strategies used by participants. We think that this explanation is unlikely because the task design limited possible strategy choices. Participants verified the

sentence with a quantifier by comparing their threshold to the proportion given as a number. Although the Approximate Number System (Dehaene, 1997) could have interfered with the precise number system, it is rather unlikely that participants were unable to precisely compare proportions. In our task, there was no time pressure on the decision and the proportions were displayed on the screen for an unlimited period of time. We feel confident in rejecting the explanations based on the variability in verification strategies as a source of observed individual differences in thresholds.

The individual differences in thresholds are also unlikely to be a result of the different cognitive abilities of our participants. We did not measure the working memory or executive function performance of participants, but our task was relatively easy and did not require much working memory or other cognitive function resources. Moreover, we included a response error parameter in our model, which accounted for variability in task performance (e.g., attention lapses or mistakes). We found that the majority of participants belonged to a low response error cluster, indicating that they performed the task on at a similar level of accuracy. Altogether, we conclude that the differences in thresholds between groups are due to different representations of the truth conditions of quantifiers.

2.4.4 Conclusions

In the current study, we identified three clusters of participants assigning different meanings to vague quantifiers such as *most*, *many*, and *few*. We showed that these quantifiers have different positions on the mental scale in subgroups of participants. Moreover, we separated individual semantic differences in meaning representations, such as vagueness and threshold, from general cognitive abilities reflected in a response error parameter. Our findings are consistent with the claim that logical words can have various semantic representations for different speakers. We believe that our approach could be helpful for studying individual differences in the representation of not only quantifiers but also other function or content words.

Chapter 3

Uncovering the structure of semantic representations using a computational model of decision-making[1]

Abstract According to logical theories of meaning, an expression's meaning can be formalized and encoded in truth conditions. The vagueness of the language and individual differences between people are a challenge to incorporate in the meaning representations. In this paper, we propose studying linguistic behavior as a decision-making process by applying computational modeling (Diffusion Decision Model) to a sentence verification task. We mapped different aspects of meaning onto model parameters. We selected two widely discussed natural language quantifiers *most* and *more than half* for a case study. We choose an experimental paradigm that allowed us to exclusively test meaning representations of quantifiers and control the pragmatic and processing differences. Our model accounts for the different types of vagueness as well as for interindividual and intraindividual differences in meaning representations. We found substantial interindividual differences in representations of *most*, along with stability of both quantifiers' representations over time. Moreover, we found that the verification process of *most* is proportion-dependent. These findings challenge semantic theories that assume the truth-conditional equivalence of *most* and *more than half*. The current study presents a promising approach to study semantic representations, which can have a wide application in experimental linguistics.

[1]This chapter is based on the manuscripts: Ramotowska, Steinert-Threlkeld, van Maanen, and Szymanik (2021). Uncovering the structure of semantic representations using a computational model of decision-making, (unpublished manuscript), and Ramotowska, Steinert-Threlkeld, van Maanen, Szymanik, (2020a). Individual differences in semantic representations: The case of *most* and *more than half* (unpublished manuscript).

3.1 Introduction

Human language is an exceptionally complex phenomenon. The meaning of words and sentences has been studied from a variety of perspectives, ranging from formal semantics descriptions through computational simulations to experimental semantics (e.g., Carcassi & Szymanik, 2021; Deschamps et al., 2015; Frank & Goodman, 2012; Hackl, 2009; Pietroski et al., 2009). Formal semantics analyzes linguistic meaning on an abstract level and studies the logical form of meaning and the composition of meanings. This approach might not immediately give testable predictions about the linguistic behavior, such as the sentence verification process, usage of the expression in the context, etc. At the same time, linguistic behavior is the product of several factors (e.g., the effect of context, vagueness, ambiguity, etc.) and individual differences in cognitive ability (e.g., lexicon size, intelligence, working memory) (Kidd et al., 2018). While semantic theories usually elaborate on the former aspect, they widely ignore the latter. Therefore, the link between the semantic theories' predictions and linguistic data is often obscured. The problem of linking between theory and data is a common problem in cognitive science (Guest & Martin, 2021). Computational modeling can be a solution because it helps to formalize our theoretical intuitions and make the theories transparent and testable.

In this paper, we argue that linguistic behavior can be studied as a decision-making process. Our approach applies to many psycholinguistic tasks including, but not limited to, lexical decision, inferences, and verification. To illustrate this idea, we apply a formal model of decision-making to data from a quantifier verification task. This task is a commonly used paradigm to test meaning representations (Bott, Augurzky, Sternefeld, & Ulrich, 2017; Clark, 1976; Deschamps et al., 2015; Just & Carpenter, 1971; Szymanik, 2016; Zajenkowski & Szymanik, 2013). In this type of task, participants have to indicate whether a sentence is true or false given an example scenario. Participants base their judgment on the provided evidence (e.g., a sentence or picture). Verification is a process of collecting evidence in favor of all decision options. Once the individual has collected sufficient evidence for one of the decision options (e.g., the true or false answer), the decision can be made. Following this observation, we show that the performance in the verification task can be analyzed using computational models of decision-making such as the Diffusion Decision Model (DDM) (Ratcliff, 1978), which is widely used in many cognitive domains outside of language (Ratcliff & McKoon, 2008; Ratcliff et al., 2016). We illustrate how this computational model allows us to disambiguate the role various aspects of meaning play in the subjects' behavior.

As a case study, we test the differences between two natural language quantifiers: *most* and *more than half*. We chose these quantifiers because they have drawn much attention from semanticists in recent years (Hackl, 2009; Lidz et al., 2011; Pietroski et al., 2009; Register et al., 2018; Solt, 2016). We show that

our modeling strategy and experimental design allow us to disentangle different aspects of meaning such as semantic representations, vagueness, pragmatics, or processing, usually confounded in experimental and corpus data. In this way, we exhibit a systematic approach to infer the meaning of quantifiers from linguistic data and to validate the semantic theories. In the next section, we will present the linguistic discussion about *most* and *more than half* and introduce the DDM.

3.1.1 Semantic representations: the case of *most* and *more than half*

Most and *more than half* – differences in processing and pragmatics

The meaning of quantifiers such as *most* and *more than half* can be expressed on the grounds of logical theories of meaning, for example, Generalized Quantifier Theory (GQT) (Barwise & Cooper, 1981; Mostowski, 1957), in the form of truth conditions. For an illustration consider the following example. The sentence "*Most/More than half* of the students passed the exam" is true under one of two conditions: (1) if the number of students, who passed the exam ($|[\![\text{Students}]\!] \cap [\![\text{passed exam}]\!]|$) is greater than half of all students ($\frac{1}{2}|[\![\text{Students}]\!]|$), or (2) if the number of students, who passed the exam is greater than the number of students, who did not pass the exam ($|[\![\text{did not pass exam}]\!]|$). Following Hackl (2009) linguistic analysis, we can formulate example truth conditions for *more than half* and *most*:

Example
(1) *More than half* of the students passed the exam. $\leftrightarrow |[\![\text{Students}]\!] \cap [\![\text{passed exam}]\!]| > \frac{1}{2}|[\![\text{Students}]\!]|$
(2) *Most* of the students passed the exam $\leftrightarrow |[\![\text{Students}]\!] \cap [\![\text{passed exam}]\!]| > |[\![\text{did not pass exam}]\!]|$

A brief reflection on these two examples leads to the conclusion that the truth conditions of *most* and *more than half* are logically equivalent. The truth value of the sentences with *most* or *more than half* will be the same in every situation. For example, both sentences will be false if 9 out of 20 students passed or true if 11 out of 20 students passed. Therefore, the logical theories of meaning predict that these quantifiers are used interchangeably.

Contrary to this claim, there is ample evidence that *most* and *more than half* lead to different linguistic behavior. These differences have been attributed to verification (Hackl, 2009) or pragmatics (Solt, 2016). Hackl (2009) postulated that both *most* and *more than half* are associated with different verification strategies, even though they are logically equivalent. The expression *most* (Example (2)) is equivalent to the expression *more than half* (Example (1)) in terms of truth conditions, but they have different linguistic representations, which yield differ-

ent verification strategies. A verification strategy is an algorithm to determine the truth value of the sentence. The algorithm to verify *more than half*(A, B) requires the computation of the intersection of sets A and B (e.g., students who passed the exam) and half of the set of all As (e.g., half of all students). The algorithm for *most*(A, B), in turn, is a vote-counting strategy, which involves tracking whether the amount of As that are B is greater than the amount of As that are not B. These two algorithms correspond to different cognitive processes and lead to different behavior. The idea that *most* and *more than half* differ in processing was further developed by associating *most* with the Approximate Number System (ANS, Lidz et al., 2011; Pietroski et al., 2009; Solt, 2016) and *more than half* with the Precise Number System (Solt, 2016).

Except for the possible processing differences, *most* and *more than half* are also associated with different pragmatics (Kotek et al., 2015). A corpus study (Solt, 2016) provided evidence that *most* is used more frequently with uncountable nouns (for example, "*most* of the sadness," Solt, 2016, p. 67) or is not suitable to use in some contexts when the proportion is close to 50% (for example, "*Most* of the American population is female" vs. "*More than half* of the American population is female," Solt, 2016, p. 67). Furthermore, Carcassi and Szymanik (2021) supported a pragmatic explanation by embedding it into the computational framework of Rational Speech Acts (Frank & Goodman, 2012).

Taken together, the linguistic puzzle about *most* and *more than half* is an excellent case, which could benefit from computational modeling. The proposed explanations of differences between *most* and *more than half* fall into two main categories: the processing explanation (e.g., differences in verification strategies), and the pragmatic explanation (e.g., the context in which *most* is used). In this paper, we argue for a third possible explanation, namely the semantic explanation: *most* and *more than half* in fact have different truth conditions. The processing and pragmatic explanations assume that *most* and *more than half* have equivalent truth conditions and that other factors cause different linguistic behavior for these quantifiers. However, none of the previous studies explicitly tested the truth-conditional equivalence of *most* and *more than half*. We show that, by using computational modeling and a specific experimental design, we can directly test the logical theories' predictions, that *most* and *more than half* have equivalent semantics, and at the same time account for other linguistic and psychological factors connected to these meanings.

Complicating factors: vagueness and individual differences

It is uncontroversial that many natural language expressions are vague. A cognitively realistic approach of semantic representations therefore has to take vagueness into account, as well as possible individual differences in semantic representations (Kidd et al., 2018). Therefore, we consider vagueness as a possible source of differences between *most* and *more than half*, and by using computational

modeling we show how it relates to differences in truth conditions.

In the domain of quantifiers, the most prominent example of vague quantifiers is *many* and *few*. Vagueness can relate to threshold (Solt, 2015), a cut-off point (e.g., proportion) for which the responses change the truth value. Solt (2011) postulated that *most* is also a vague quantifier with a fuzzy threshold, while *more than half* has sharp meaning boundaries and a 50% threshold. The vague quantifiers are sensitive to individual differences in thresholds (Ramotowska et al., 2020b). Specifically, context-dependent quantifiers, such as *many* or *few*, have varying usage depending on the speaker (Yildirim et al., 2016). Moreover, they meanings are flexible and can be adjusted to another speaker's usage (Yildirim et al., 2016), learning criterion (Heim et al., 2015), and can be changed during the adaptation process (Heim et al., 2020).

In the computational model, we adopt two types of vagueness taken from the semantic categorization literature: criteria and degree vagueness (Verheyen, Droeshout, & Storms, 2019), to express the vagueness of quantifiers' meanings. Criteria vagueness is a disagreement between participants about the criteria or conditions to classify a given concept into a given category (Verheyen, White, & Égré, 2019). Degree vagueness, in turn, means that individuals agree about the classification criterion but disagree to what extent the given concept satisfies this criterion (Verheyen, White, & Égré, 2019). For example, one can believe that an abstract painting cannot be considered a piece of art because art should imitate reality (criteria vagueness). In contrast, this person can consider two realistic paintings to be art, but to a different extent depending on how well they imitate reality (degree vagueness). Both types of vagueness can cause between-participants variation in semantic categorization (Verheyen & Storms, 2013, 2018). In addition to between-participants variability, studies also show a within-participant inconsistency in the classification of items that are not typical category members compared to items that are typical category members or items irrelevant to a category (McCloskey & Glucksberg, 1978). Moreover, the categorization criterion can change within one month's time (Verheyen, White, & Égré, 2019) or during a more extended period, for example, because of aging (Verheyen, Droeshout, & Storms, 2019).

To summarize, the logical theories give a clear prediction about participants' behavior concerning *most* and *more than half*. Participants should apply the same criteria (truth conditions) to classify the given proportion as *most* or *more than half*. Therefore, there should not be any variation between participants nor within participants. On the other hand, if the meaning of *most* is vague, we could expect the between-participants or within-participants variation in representations of *most*. More specifically, if *most* is a vague quantifier in the criterial sense, we can expect that participants to differ in thresholds for *most*, and also, to possibly change their truth conditions over time. *Most* can also be vague in the degree sense. Participants can have the same truth conditions for *most*, but they can differ in the way they found given proportions suitable to use with *most*. This

would lead to observed uncertainty around thresholds, rather than a shift in truth conditions. We operationalized uncertainty as longer reaction times around the threshold.

3.1.2 Modeling, experiments, and predictions

In the previous section, we identified possible sources of differences between *most* and *more than half*. We argued that the linguistic discussion around these two quantifiers cannot be decided based on existing experimental and corpus data. The goal of the current paper is to understand linguistic behavior in a decision-making framework. For our purposes, we chose the Diffusion Decision Model (DDM).

Diffusion Decision Model

The DDM (Ratcliff, 1978) is a canonical evidence accumulation model that represents two-choice decisions (e.g., between true and false) as a noisy evidence accumulation process toward two decision boundaries (Figure 3.1). The model assumes that a decision is reached as soon as one of the boundaries is crossed, with the time required to reach the boundary called the decision time. In this way, both reaction times and the proportion of decision outcomes can be jointly investigated. The parameters that specify the DDM are typically found to match specific cognitive processing components (Mulder, van Maanen, & Forstmann, 2014). In this way, various sources of variability of behavior can be separated. The parameter that expresses the speed of the accumulation process is called drift-rate (v). The evidence accumulation process starts at one point (z) and finishes when enough evidence is accumulated toward one of the decision options. The decision options are operationalized as two separate decision boundaries (separated by a). Moreover, DDM assumes that the reaction times during the decision-making process consist of a decision time and a non-decision time (*Ter*, the time required to execute the response after the decision is made). Additionally, the model allows for the trial-by-trial variability in the starting point (s_z), drift rate (s_v) and non-decision time (s_{Ter}).

Because the DDM accounts for both responses and reaction times distributions, it makes it possible to extract richer information from the data than the traditional comparison of mean reaction times or accuracies between groups or individuals. It also allows for analyzing the relationship between accuracy and reaction time distributions rather than treating them as two separate measures.

The DDM was successfully applied to model cognitive processes in two domains that are relevant for the current study: the domain of number cognition (Ratcliff & McKoon, 2018, 2020; Ratcliff, Thompson, & McKoon, 2015) and the domain of individual differences in linguistics tasks (Pexman & Yap, 2018; Ratcliff et al., 2010; Yap, Balota, Sibley, & Ratcliff, 2012). For example, Ratcliff et

al. (2010) found that participants with a lower IQ score had lower drift rates in a lexical decision task than participants with a higher IQ. Pexman and Yap (2018) fitted DDM to data from a categorizations task (concrete vs. abstract words) to test the individual differences in semantic processing and decision-making. They found differences in drift rates between participants with high vs. low vocabulary knowledge. Yap et al. (2012) found an association between DDM parameters (drift rate, boundaries separation, and non-decision time), performance in a lexical decision task, and individual differences in vocabulary knowledge. In the domain of number cognition, Ratcliff et al. (2015) showed that individual differences in accuracy or reaction times in numeracy tasks are explained by different parameters.

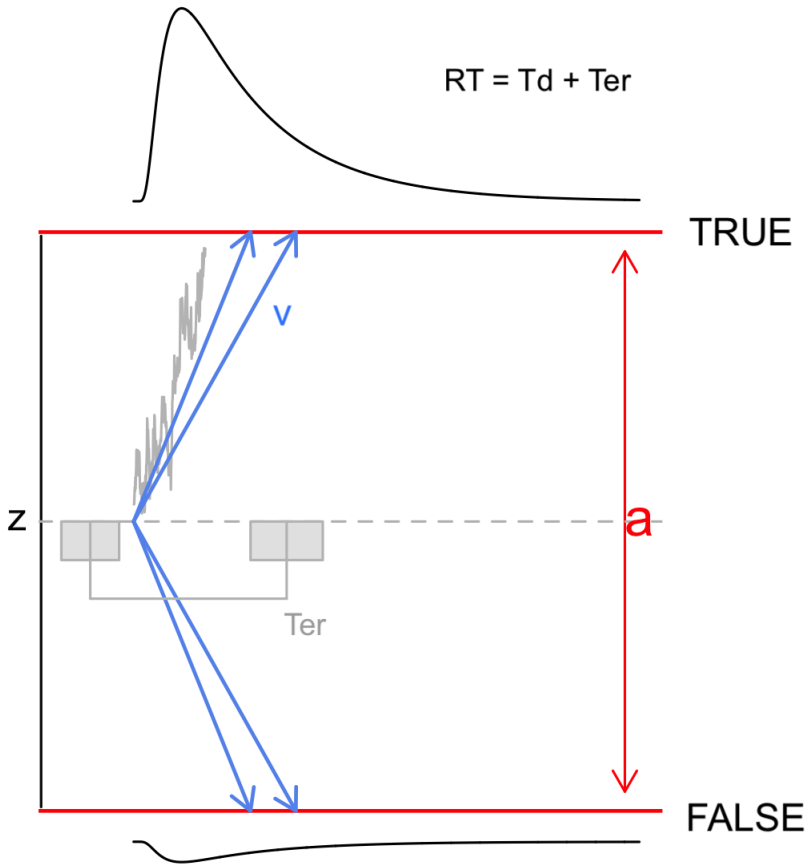


Figure 3.1: Representation of the Diffusion Decision Model. The accumulation process starts at point z (starting point) and finishes when one of the decision boundaries (TRUE/FALSE response) is reached (decision boundaries separation parameter a). *Ter* represents non-decision time; v represents drift rate. Reaction times (RT) are the sum of non-decision time (*Ter*) and decision time (*Td*).

Taken together, evidence accumulation models such as DDM seem to be an excellent tool to analyze response proportion and reaction time distributions data in linguistic and number cognition tasks. Moreover, the analysis of the participants' parameters gives a meaningful interpretation of individual differences between subjects. Finally, the flexibility of the evidence accumulation model provides the opportunity to adapt them to different tasks. For these reasons, the DDM also seems ideally suited to the analysis of semantic data, as we will illustrate below.

Experimental design

To test the semantic representations of *most* and *more than half*, we chose a specific experimental paradigm in which we limited pragmatic and processing differences between these quantifiers. In this way, we were able to test the equivalence of the truth conditions for *most* and *more than half*. In our task, participants ver-

ified a sentence with a quantifier of the form "Q of the As are B," where Q was one of the quantifiers: *most*, *more than half*, *fewer than half*, *few* and *many*, against the sentence with a proportion ("X% of the As are B") given as a number e.g., 55%. We included *fewer than half*, *few* and *many* for control and generalization purposes.

To ensure that we studied the semantic properties of *most* and *more than half*, we made two important choices regarding experimental design. Firstly, we decided to eliminate the potential differences in processing and verification strategies between the two quantifiers. We chose to present proportion as a number and we did not limit the time to make a decision in order to enforce the precise processing of the numerosities. In this way, the results obtained cannot be attributed to the usage of the ANS (Dehaene, 1997) during the verification of *most* compared to *more than half*. We also used pseudowords for As and Bs to limit pragmatic inferences (van Heuven et al., 2014). Therefore, our participants could only access the meaning of quantifier and proportion. In Experiment 1, we tested whether the variation of individual thresholds differed between quantifiers. In Experiment 2, we tested the stability of individual thresholds over time.

Predictions

We ran two experiments testing the semantic representations of *most* and *more than half* (see the direct replication of in Appendix B) and one experiment testing the stability of these representations over a two-week period. We tested predictions that follow directly from logical theories (LT). **LT Hypothesis**: *Most* will have the same threshold (50%) as *more than half* for all participants and the threshold for both quantifiers will be stable over time. Moreover, we considered the effect of different types of vagueness. Following the hypothesis that the meaning of *most* is vague (Solt, 2011), we formulated a complementary, criterial vagueness (CV) hypothesis. **CV Hypothesis**: *Most* will have a greater variation in individual thresholds than *more than half*, meaning that some participants will have a higher threshold for *most* than *more than half* and some participants will have a 50% threshold for both quantifiers. With regard to the stability of thresholds over time, we considered that *more than half* should have a stable threshold and we could observe the variation in the thresholds for *most* over time. We also made a prediction about the degree vagueness (DV) of *most* (uncertainty about the response reflected in longer reaction times). **DV hypothesis**: The speed of verification of *most*, but not *more than half*, will be proportion-dependent, meaning that reaction times for proportions close to the individual threshold will be longer than for proportions further from the threshold. We also predicted that the proportion-dependent relationship will be reflected in modeling data by degree vagueness parameter. Moreover, we tested whether the two types of vagueness can interact. We considered the logical possibility that (1) *most* and *more than half* can have the same 50% threshold, and yet *most* can have a higher degree

vagueness (longer reaction times around 50% proportion); (2) *most* can have greater variation in thresholds and a higher degree of vagueness than *more than half* around the individual threshold; (3) or *most* can have greater variation in thresholds and the degree vagueness interacts with the choice of threshold. This analysis was exploratory in nature.

In addition, we looked into other model parameters. We predicted that the vagueness of *most*, *many* and *few* should be reflected in model parameters. Following Schlotterbeck et al. (2020), we expected to replicate differences in non-decision time and drift rate for negative vs. positive quantifiers (see Appendix B).

3.2 Experiment 1

3.2.1 Methods

Participants

We tested 90 users of the Amazon Mechanical Turk platform (`https://www.mturk.com/`). We included 72 participants (48 male) in our analysis, age: $M = 35$, $SD = 11$, range: 22–59. The sample represented various educational backgrounds: high school graduates (24 participants), high school graduates, who started college (22 participants), and college graduates (26 participants). The subjects received USD 4 in compensation for their participation. The study was approved by the Ethics Committee of the University of Amsterdam's Faculty of Humanities.

Exclusion criteria

We applied two exclusion criteria. Firstly, we excluded fast guessing participants (11 subjects), whose reaction times were faster than 300 ms for 50% or more responses. Additionally, we tested whether participants respected quantifier monotonicity, i.e. for the quantifiers *most*, *more than half*, and *many* we expected the probability of saying 'true' to increase with the increase of the proportion. The opposite effect was expected for *few* and *fewer than half*. We tested this assumption by estimating random slopes for proportion for each participant (*glmer* function in R package *lmerTest*, Kuznetsova et al., 2017). For positive quantifiers, we excluded participants with negative slopes, and for negative quantifiers, we excluded those with positive slopes. Based on this criterion, we excluded together 6 participants. Finally, we excluded 1 participant who had participated in a similar experiment previously.

Design

We decided to use pseudowords in order not to introduce any other variability in meaning beyond different quantifiers. We used *Wuggy* software (Keuleers & Brysbaert, 2010) to generate pseudowords from English nouns (As) and adjectives (Bs). We selected 50 pseudo-adjectives and 50 pseudo-nouns, which a native English speaker indicated sounded like plausible English words. We controlled for the frequency of the original English words. The words on both final lists had a *Zipf*-value of 4.06 (SUBRLEX-US database, van Heuven et al., 2014). All words on the final lists were six-letter words. We matched each quantifier with each pair of pseudowords. We presented the trials in random order.

Procedure

Participants saw two sentences on separate screens. On the first screen, they saw a sentence with a quantifier and pseudowords A and B: "*most/more than half/many/ few/ fewer than half* of the As are B." To display the first sentence, participants had to press the down arrow key and keep it pressed as long as they wanted to read the sentence. When they released the down arrow key, the sentence disappeared. To display the second sentence, they had to press the down arrow key again (but they did not have to keep it pressed). On the second screen, participants saw a sentence with a proportion given as a percentage: "p% of the As are B," where As and B were the same pseudowords as in the first sentence. The proportion, p%, in the second sentence was randomly drawn from 1% to 99%, excluding 50%. On this screen, participants had to decide whether the first sentence was true or false based on the information from the second sentence by pressing the left arrow or right arrow key (counterbalance between-participants). We presented first the sentence with a quantifier and second the sentence with a proportion, because quantifiers had different lengths. This factor could affect the reading times of sentences with quantifiers and therefore, the reaction times and the estimation of DDM parameters.

We counterbalanced within-participant proportions above and below 50% for *more than half*, *fewer than half*, and *most*. Altogether, participants saw 250 trials, 50 mixed trials for each quantifier. The experiment was preceded by a short training block (8 trials). In the training block, participants saw sentences with the quantifiers *some*, *all*, and *none*. At the end of the experiment, participants filled in a short demographic survey.

Preprocessing reaction times (RT) data

Apart from excluding participants, we also excluded reaction times that were too short or too long. Before we fitted the DDM, we excluded reaction times shorter than 300 ms and longer than mean+2SD for each quantifier and each response type (true, false) separately.

3.2.2 Computational modeling of the verification process

Drift Diffusion Model - Drift rate

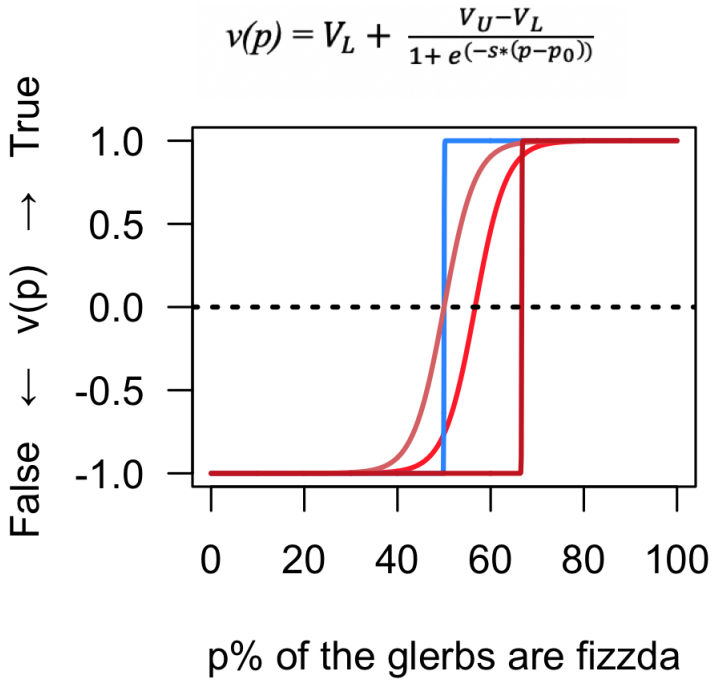


Figure 3.2: Drift rate structure, where V_L is the lower asymptote, V_U is the upper asymptote, s is the growth rate, p is the proportion in the second sentence of each trial, and p_0 is the individual threshold. The figure presents the predicted drift rate *v(p)* for *more than half* (blue line) and three possible drift rates *v(p)* for *most*: with threshold at 50%, but greater DV (the lightest red line), with threshold higher than 50% and the same DV as *more than half* (the darkest red line) and with threshold higher than 50% and greater DV (middle red line).

We fitted the simple DDM (without variability parameters s_z, s_v, and s_{Ter}) to reaction times and responses data for each participant. We did not include variability parameters in order to simplify the model and because we did not expect them to differ between quantifiers. We assumed that the differences in threshold (midpoint parameter p_0) and degree vagueness (growth rate parameter s) between *most* and *more than half* would manifest during the verification process in the drift rate. We operationalized these concepts via the generalized logistic function presented in Figure 3.2. The p_0 was the proportion for which the drift rate was zero and the s indicated the steepness of the drift rate (the higher the s the steeper the drift rates).

Bayesian model averaging

Because one of the goals of our study was to capture individual differences between participants, we considered that there might also be individual differences between participants in terms of which model is best according to Akaike Information Criterion (AIC) values (Akaike, 1998). Therefore, we decided to use Bayesian model averaging (BMA) for all DDM parameters (Hoeting, Madigan, Raftery, & Volinsky, 1999; Miletić & van Maanen, 2019; Wagenmakers & Farrell, 2004) rather than parameters from the winning model. BMA is a method to compute parameters for each participant, taking into account the weighted average of the parameters from each model.

The weight for model i (w_iAIC) is defined using the AIC values (Wagenmakers & Farrell, 2004):

$$w_iAIC = \frac{e^{-\frac{1}{2}\Delta_i(AIC)}}{\sum_{k=1}^{K} e^{-\frac{1}{2}\Delta_k(AIC)}} \tag{3.1}$$

Where $\Delta_i(AIC) = AIC_i - min(AIC)$ for each model i.

Mixed-effects regression modeling

For all linear mixed-effects regression models, we applied the individual BMA thresholds to each participant's response data from Experiment 1. For positive quantifiers, we included false responses below the threshold and true responses above the threshold in the analyses (inversely for negative quantifiers). To test the DV hypothesis, we fitted a linear mixed-effects model (*lmer* function in *lmerTest* package in R, Kuznetsova et al., 2017) with log-transformed (log10) reaction times (in seconds) as the dependent variable and z-scored proportion, quantifier (*most*, *more than half*), response (true/ false), and their interactions as predictors. We used the response true and the quantifier *most* as the baseline. For exploratory analysis, we fitted a linear mixed-effect models (*lmer* function in *lmerTest* package in R, Kuznetsova et al., 2017) with log-transformed (log10) reaction times (in seconds) as a dependent variable and distance from threshold (proportion centered on individual threshold), individual threshold centered on mean threshold, response (true or false), and their interactions as predictors for each quantifier. We used the response true as the baseline. We used log-transformation of reaction times to improve the compatibility of mixed-effects models with their assumptions: normality and homoscedasticity of residuals.

We applied the same procedure of testing the random effect structure for all models. We tried to keep the random structure of the model maximal (Barr, Levy, Scheepers, & Tily, 2013). Therefore, we always included by-subject random intercept and by-subject random slopes if they improved the model (determined by *anova* function in R). We included by-item random intercept if the model was not overfitted.

3.2.3 Results

Descriptive statistics

Table 3.1 summarizes the mean reaction times and proportion of true and false responses for each quantifier. This summary already suggests the differences between *most* and *more than half*. Firstly, the reaction times for *more than half* are shorter than for *most*. Secondly, *most* has a greater proportion of false than true responses, indicating the possible difference in threshold. We applied DDM to further explain these effects.

Table 3.1: Mean reaction times (RT) in seconds (*SD*) and proportion of response true vs. false in Experiment 1.

Quantifier	Response true		Response false	
	RT	response	RT	response
Few	1.201 (.133)	.39	1.087 (.134)	.61
Fewer than half	1.170 (.186)	.48	1.064 (.115)	.52
Many	1.000 (.117)	.57	1.107 (.122)	.43
Most	1.044 (.395)	.47	1.038 (.160)	.53
More than half	.917 (.086)	.50	.942 (.092)	.50

Model fit and comparison

We used R package *rtdists* to fit the DDM, and we estimated the maximum likelihood of DDM parameters using particle swarm optimization (Clerc, 2010), on the seconds scale of reaction times. To identify model parameters, we fixed the diffusion coefficient, which indicates the standard deviation of the random noise in the diffusion process, to a scaling constant of 0.1.

To test which DDM parameters were identical across conditions, we systematically varied the model constraints. We evaluated each constrained model by assessing AIC values (Akaike, 1998). We used the AIC values to compute the number of participants for whom each model was the best model (n best) or one of the three best models (n top 3). Based on individual AIC values of each participant, we assigned ranks to each model (1 for the best model; 9 for the worst model). Next, we computed the *mean* rank for each model (the lower the rank, the better the model). Model comparison descriptive statistics are summarized in Table 3.2.

Table 3.2: Model comparison in Experiment 1 (M. is Model, k is the number of free parameters in the model; *mean* is the mean rank; n best is the number of participants for whom the given model was the best; n top 3 is the number of participants for whom the given model was one of the three best models).

M.	Parameters			Rank		
	Free	Fixed	k	*mean*	n best	n top 3
1	*Ter*, *a*, *z*, p_0, *s*, V_L, V_U		35	7.72	1	4
2	*a*, *z*, p_0, *s*, V_L, V_U	*Ter*	32	6.60	3	11
3	*z*, p_0, *s*, V_L, V_U	*Ter*, *a*	28	5.53	4	16
4	p_0, *s*, V_L, V_U	*Ter*, *a*, *z*	25	5.10	4	22
5	p_0, V_L, V_U	*Ter*, *a*, *z*, *s*	22	3.99	14	30
6	p_0	*Ter*, *a*, *z*, *s*, V_L, V_U	16	4.14	10	31
7	p_0	*Ter*, *a*, *z*, *s*, V_L, V_U	14	4.39	3	27
8		*Ter*, *a*, *z*, *s*, V_L, V_U, p_0	12	**3.38**	16	**41**
9		*Ter*, *a*, *z*, *s*, V_L, V_U, p_0	11	4.17	**17**	34

For some parameters (*Ter*, *z*, V_L and V_U), we observed differences between positive vs. negative quantifiers or, in case of parameter *s*, between *more than half/fewer than half* and the rest of quantifiers. We started the parameter estimation process with an unconstrained model in which all parameters could differ for all quantifiers (Model 1). We then chose *Ter* to be the same across negative quantifiers (*few* and *fewer than half*) and positive quantifiers (*many*, *most*, *more than half*) (Model 2). In Model 3, we constrained *a* parameter to be the same across all quantifiers. The *z* parameter was the same across negative quantifiers (*few* and *fewer than half*) and positive quantifiers (*many*, *most*, *more than half*) (Model 4). We constrained parameter *s* to be the same for *fewer than half* and *more than half* and the same for *most*, *many*, and *few* (Model 5). We also constrained asymptotes parameters V_L and V_U (Model 6), in the same way as the *Ter* and *z* parameters. Finally, we tested the model with symmetric V_L and V_U parameters (Model 7). In Model 8 we constrained p_0 parameters for *more than half* and *fewer than half* and in Model 9 also for *most*.

In addition to model comparison, we also visually investigated the individual participant model fit and the aggregate Model 7 fit (Figure 3.3) by means of the Vincentizing method (Ratcliff, 1979). Figure 3.3 presents a cumulative probability of reaction time data separately for true and false responses. Each percentile of the data was scaled by the overall proportion of the true or false responses. The cumulative distribution shows the proportion of true or false responses for each percentile. For example, the true responses for *few* make up 39% of all responses (see Table 3.1) and the 0.95 percentile of true responses has a cumulative probability of 0.37 as it covers 95% of true responses.

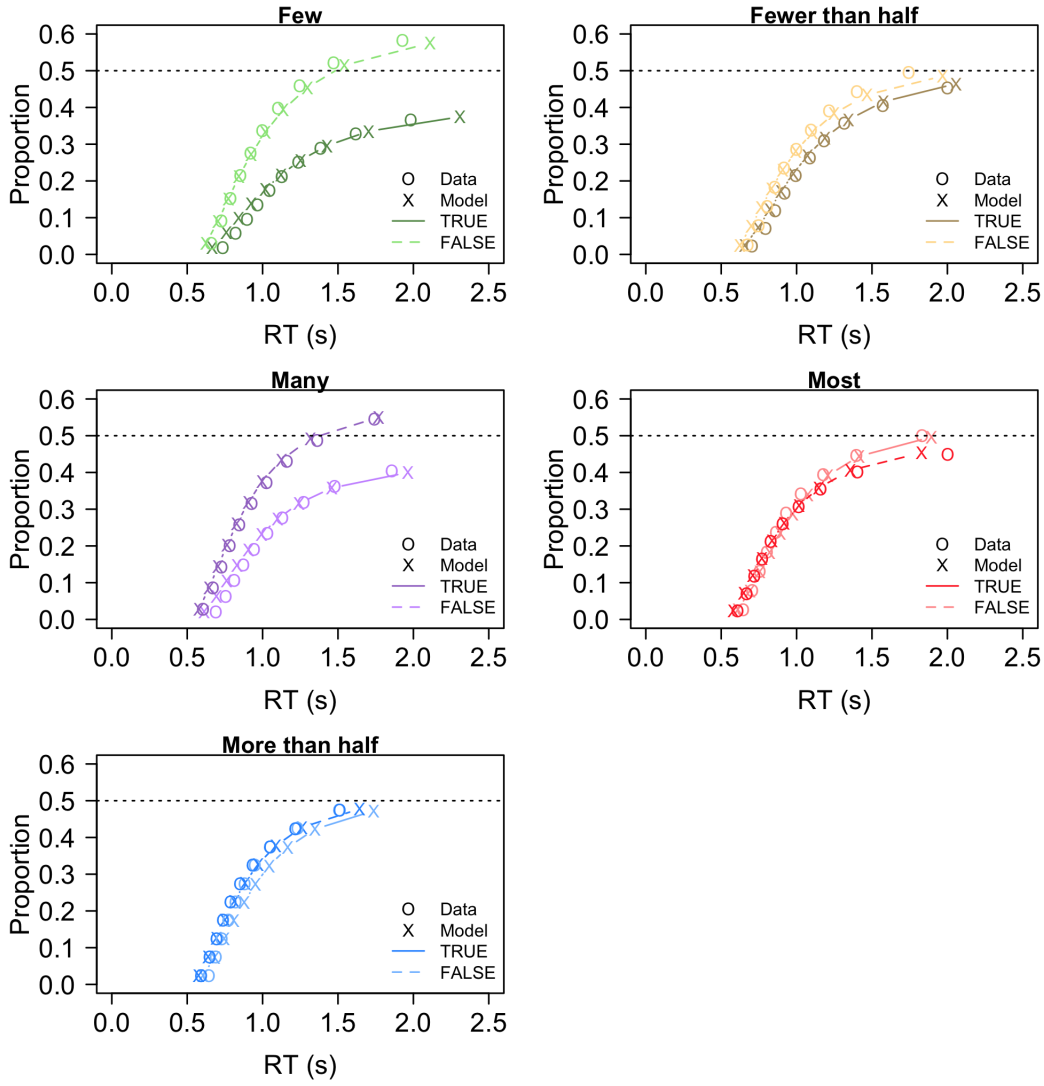


Figure 3.3: Defective cumulative density plots show the average fit of Model 7. For this visualization, we plot the mean 5%, 15%, 25%, 35%, 45%, 55%, 65%, 75%, 85%, and 95% percentiles over participants, scaled by the proportion of true and false responses, separately for the data and the Model 7 prediction.

Finally, we used AIC values to compute weights for each model and weighted averaged parameters (BMA, Hoeting et al., 1999, Table 3.3 and Figure 3.5). The large variation in best model fit (see Table 3.2) supports our choice to use BMA parameters. Figure 3.4 shows the example drift rates for Model 7. The drift rates for *more than half* and *fewer than half* are steeper and have lower variability in thresholds (proportion for which $v(p) = 0$) than drift rates of other quantifiers.

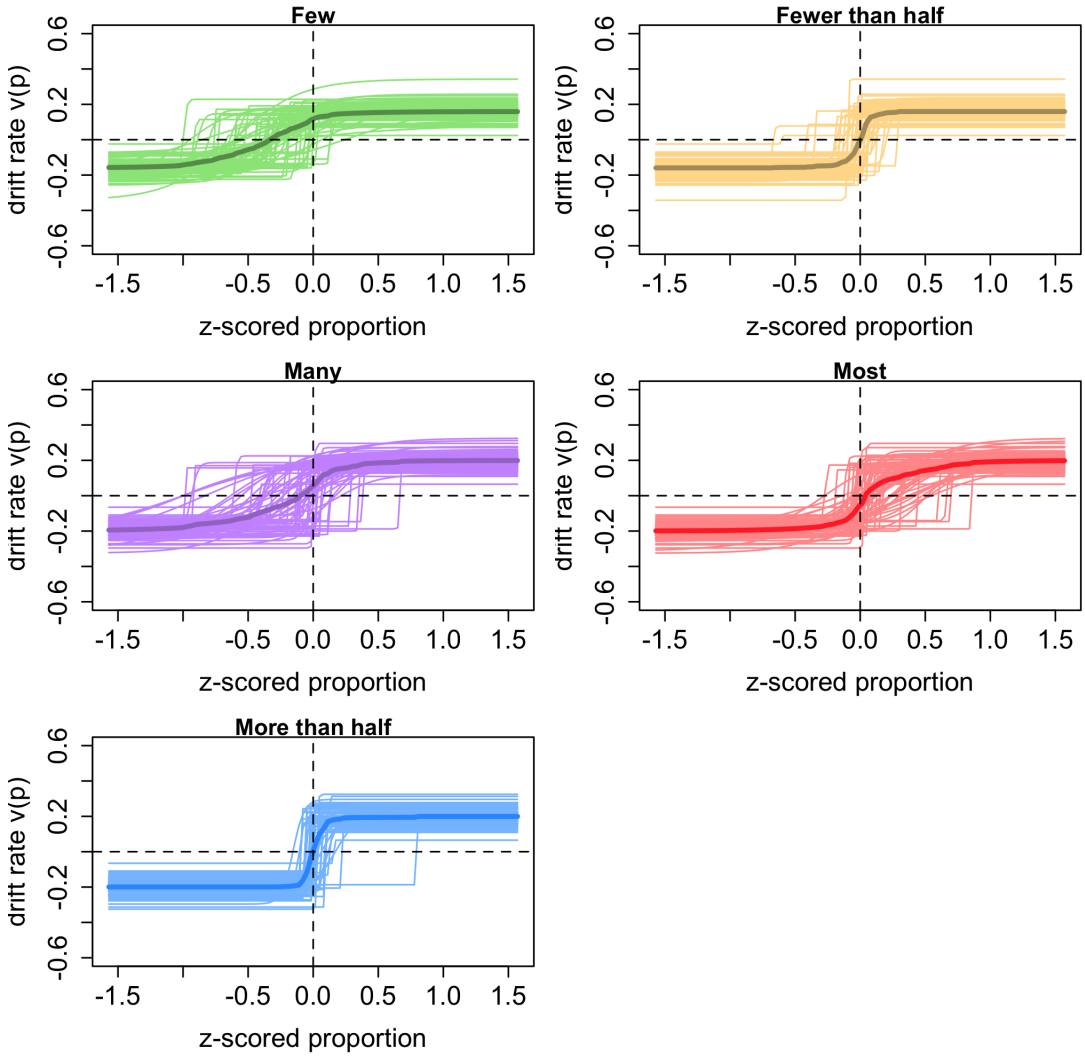


Figure 3.4: Drift rates in Experiment 1 for Model 7. Darker lines indicate mean drift rate and lighter lines indicate individual drift rates. Individual participants threshold p_0 is the proportion on the x-axis for which drift rate $v(p)$ on the y-axis is equal to zero. The scale parameter s indicates steepness of the drift rate function.

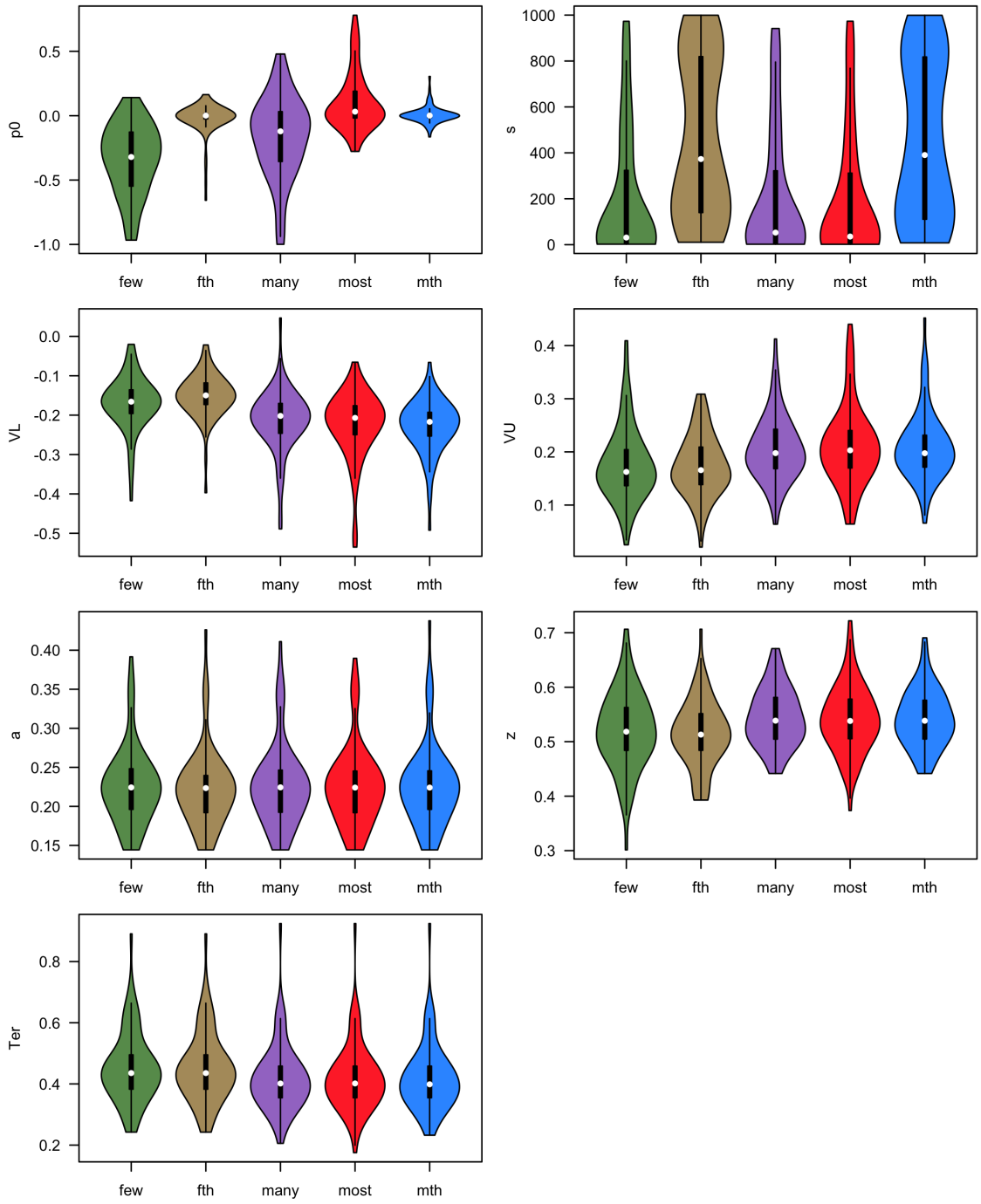


Figure 3.5: Violin plots representing the distributions of BMA parameters (Model 1 to Model 9) for all five quantifiers (fth is *fewer than half* and mth is *more than half*).

Table 3.3: Summary for model mean (*SD*) parameters after BMA (Model 1 to Model 9) using AIC.

Quantifier	p_0	s	V_L	V_U	a	z	Ter
Few	-.34	197	-.17	.18	.23	.52	.45
	(.27)	(283)	(.07)	(.07)	(.05)	(.07)	(.12)
Fewer than half	-.02	461	-.15	.17	.22	.51	.45
	(.12)	(345)	(.06)	(.06)	(.05)	(.06)	(.12)
Many	-.17	203	-.21	.21	.23	.54	.42
	(.32)	(286)	(.08)	(.06)	(.05)	(.05)	(.11)
Most	.10	196	-.23	.21	.23	.54	.42
	(.22)	(280)	(.09)	(.07)	(.05)	(.06)	(.11)
More than half	.006	456	-.23	.21	.23	.54	.42
	(.06)	(346)	(.07)	(.06)	(.05)	(.05)	(.11)

Variation in thresholds

We systematically constrained model parameters (see Table 3.2) and we compared the AIC (Hoeting et al., 1999) to determine which model had a better balance between goodness of fit and model complexity (Pitt & Myung, 2002). In order to test the LT and CV hypotheses, we constrained parameter $p_0 = 50\%$ for *more than half*, *fewer than half*, and *most* and compared this model to the model with p_0 as a free parameter (Model 7, see Table 3.2). The LT hypothesis predicts that the constrained model will be preferred for all quantifiers, while the CV hypothesis predicts this only for *more than half* and *fewer than half*.

In the first step, we constrained p_0 for *more than half* and *fewer than half*. We found that the constrained model was preferred over the model with p_0 as a free parameter by 57 out of 72 participants. Next, we constrained the p_0 parameter for *most* as well. The model fitted better for 41 out of 72 participants. These results therefore support the CV hypothesis: only some participants had a 50% threshold for *most*, and the variation in thresholds was higher for *most* than *more than half*. This difference can be observed in Figure 3.4, which shows a greater variability for *most* than *more than half* in drift rates for the model with p_0 as a free parameter (Model 7).

Accuracy with respect to threshold

Secondly, we applied the individual BMA thresholds to each participant's response data and included the true responses above the threshold and false responses below the threshold in the analysis for *most*, *many* and *more than half* (opposite for *few* and *fewer than half*). We checked the accuracy of participants relative to their threshold (Table 3.4). The overall accuracy for all quantifiers was high. *More than half* had the highest accuracy (98%) and *few* and *fewer than*

half for true responses had the lowest accuracy (90%).

Table 3.4: Relative to threshold mean accuracy (*SD*) in Experiment 1.

Quantifier	Response true	Response false
Few	.90 (.11)	.95 (.07)
Fewer than half	.90 (.11)	.95 (.08)
Many	.96 (.07)	.93 (.11)
Most	.95 (.07)	.96 (.05)
More than half	.98 (.04)	.98 (.04)

Degree vagueness

Next, we tested the DV hypothesis (see model summary in Table 3.5 and the regression models comparison in Appendix B). As expected, we found a significant effect of proportion (β = -.05; t = -6.76; $p < .001$) and a significant proportion-quantifiers interaction (β = .04; t = 3.80; $p < .001$). In addition, we found that *most* was verified more slowly than *more than half* (β = -.06; t = -5.66; $p < .001$). This finding supports the DV hypothesis that verification of *most* is proportion-dependent. The relationship between proportion and reaction times is illustrated in Figure 3.6.

Table 3.5: Summary of the model testing DV hypothesis in Experiment 1.

Effect	Estimate	t value	p value
intercept	-.0002	-.01	.99
prop	-.05	-6.76	$< .001$
quant	-.06	-5.66	$< .001$
resp	.01	1.21	.23
prop:quant	.04	3.80	$< .001$
prop:resp	.11	10.42	$< .001$
quant:resp	.02	1.12	.26
prop:quant:resp	-.07	-4.69	$< .001$

prop = proportion; quant = quantifier; resp = response

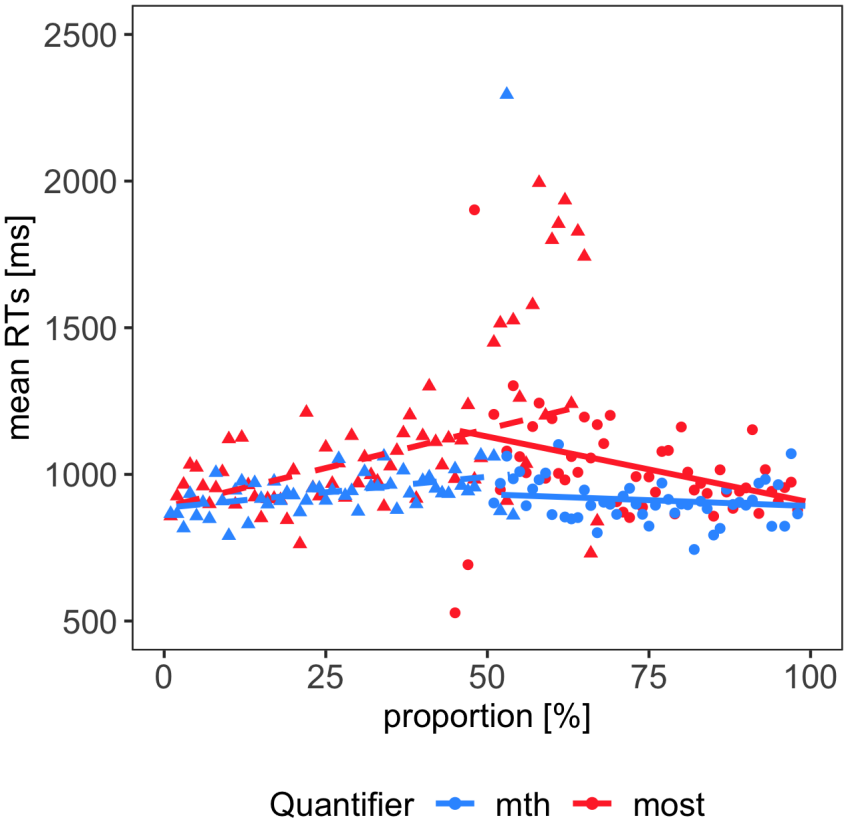


Figure 3.6: Response times as a function of proportion for *most* vs. *more than half* (mth). Each triangle represents mean reaction times (mean RTs) for proportions below the individual threshold and each dot represents mean reaction times for proportions above the individual threshold. Dashed lines represent regression lines for false responses below the threshold and solid lines for true responses above the threshold.

Effect of threshold on reaction times

Finally, we tested whether the choice of the individual threshold affects the speed of the verification process. This analysis was exploratory in nature. We expected the selection of the threshold to affect the verification process of quantifiers with various possible representations such as *many*, *few*, and *most*. We fitted a linear mixed-effects model to data from each quantifier separately.

More than half and fewer than half We tested the effect of the individual threshold on reaction times (see Table 3.6). For both quantifiers, we found that the effect of the threshold was not significant: *more than half* (β = -.009; t = -1.45; p = .15), *fewer than half* (β = .003; t = .77; p = .44). As predicted, we did not find evidence that the choice of threshold affects the speed of verification in *more than half* and *fewer than half*.

Most For *most*, we found significant effect of threshold (β = .006; t = 2.32 $p < .05$) and threshold-distance interaction (β = -.0002; t = -3.03; $p < .01$) (see Table 3.6). As expected, the verification of *most* was affected by the choice of threshold.

Many and few For *many*, we found that the effect of the threshold (β = -.003; t = -1.93; p = .056) was close to significance level 0.05, and the threshold-distance interaction (β = -.0001; t = -0.25; p = .80) was insignificant (see Table 3.6). For *few*, we found that the effect of threshold (β = -.004; t = -1.90; p = .06) was close to significance level 0.05, and the threshold-distance interaction (β = -.0002; t = -2.75; $p < .01$) was significant (see Table 3.6).

Table 3.6: Summary of the models estimates testing the effect of threshold.

Effect	*More than half*	*Fewer than half*	*Most*	*Few*	*Many*
intercept	-.06***	.02.	.003	.10***	.003
dist	-.0004.	-.00003	-.002***	.004***	-.002***
thr	-.009	.003	.006*	-.004.	-.003.
resp	.03**	-.006	.02	-.05***	.04**
dist:thr			-.0002**	-.0002**	-.00001
dist:resp	.001***	-.001**	.004***	-.005***	.004***
thr:resp	.009***		-.002	.0001	-.0006
dist:thr:resp			.0002**	.0002**	-.0001*

dist = distance; thr = threshold; resp = response;
*** $p < .001$; ** $p < .01$; * $p < .05$.

Altogether we found that the effect of the threshold was significant for *most*, and was close to significance level of 0.05 for *few* and *many*. We did not find this effect for quantifiers with sharp meaning boundaries. This finding gives a moderate support to the hypothesis that the speed of the verification depends on the choice of threshold.

3.3 Experiment 2

3.3.1 Methods

Participants

We recruited 89 participants via the Prolific platform (`https://www.prolific.co/`), and 72 participants completed both sessions. We included 64 participants (46 female) in the final sample. Participants were 32 years old on average (SD = 9, range: 18—60) and represented the following educational backgrounds: attending high school or high school graduates (5 participants), high school graduates, who

started college (6 participants), and college graduates (53 female). Participants were paid 7.5£ per hour. The study was approved by the Ethics Committee of the Faculty of Social and Behavioral Sciences of the University of Amsterdam.

Exclusion criteria

We used the same exclusion criterion as in Experiment 1. We excluded participants if they met one of the criterions in at least one testing session. We excluded 3 fast guessing participants and 4 participants who failed to meet the monotonicity criterion. In addition, we excluded 1 participant who was not a native English speaker.

Design

We used the same design as in Experiment 1.

Procedure

The stability experiment had the same number of trials as Experiment 1, and it was also preceded by a short training block. After completing the experiment participants filled in a brief demographic survey. We simplified the procedure from Experiment 1. Participants had to press the K key on their keyboard to move to the first screen, containing a sentence with a quantifier, but they did not have to keep the K key pressed. To move to the second screen, containing a sentence with proportion, participants had to press the K key again. To provide a response (true or false) they had to press the J or L key (counterbalanced across participants). Participants did the experiment twice. The proportions presented in the second sentence and the order of the trials were randomized across sessions. The second session was held in the third week after the first session.

Preprocessing reaction time (RT) data

We used the same preprocessing procedure as in Experiment 1.

3.3.2 Computational model

We applied Model 7 from Experiment 1 to data from both sessions (Table 7 shows the mean parameters). We chose this model because it included all the constraints, which highlighted differences between positive and negative quantifiers and between vague and quantifiers with sharp meaning boundaries, but had a free threshold parameter.

3.3.3 Results

Descriptive statistics

We summarize the descriptive statistics in both sessions (Table 3.7). We observed similar patterns to those observed in Experiment 1: we found shorter RTs for *more than half* than *most* and larger proportion of false responses for *most*. We also noticed that participants were faster in the second session, which indicates a learning effect.

Table 3.7: Mean RTs in seconds (*SD*) and proportion of true vs. false response in each session.

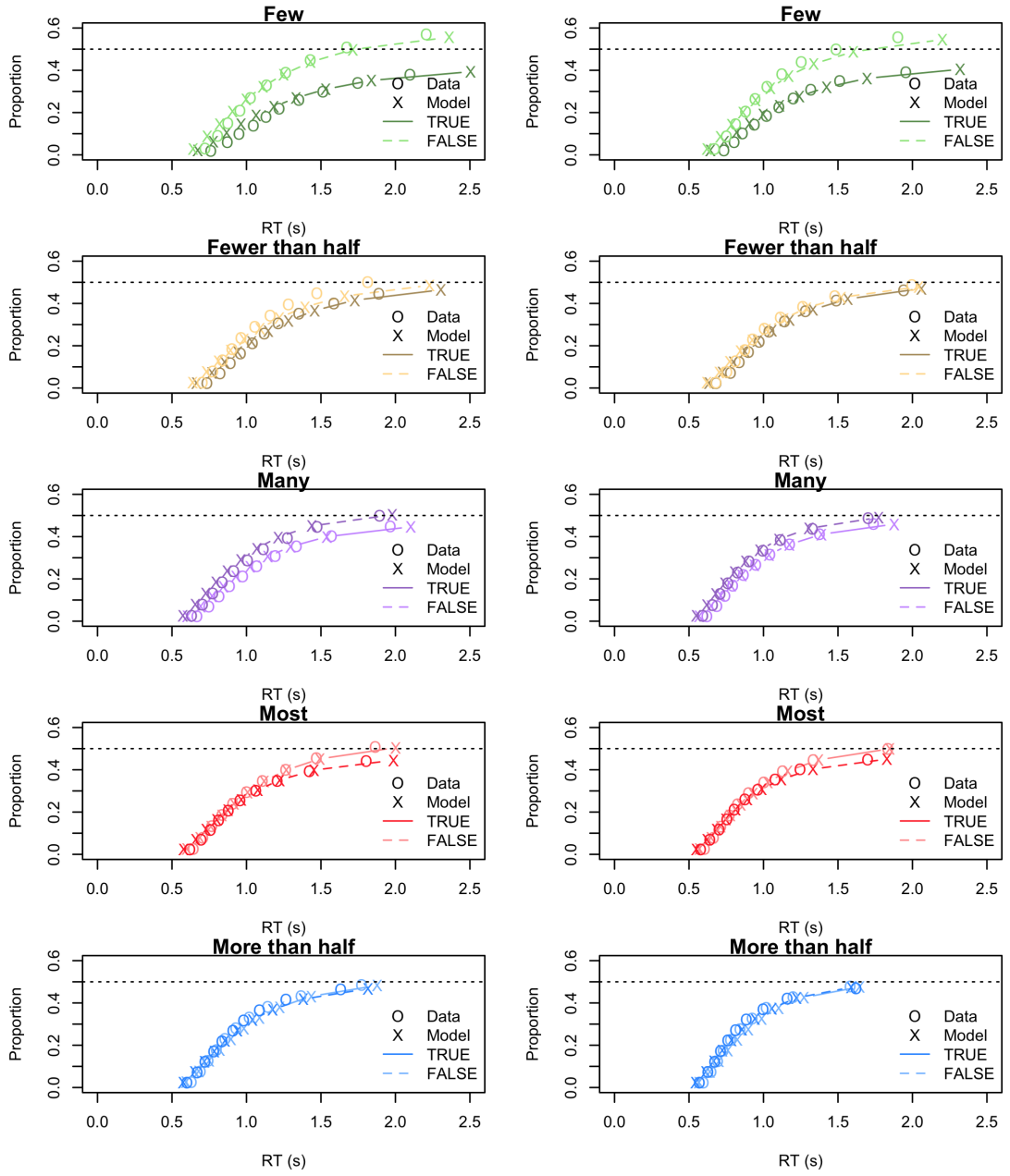
Quantifier	Truth	Session 1		Session 2	
	value	RTs	responses	RTs	responses
Few	false	1.232 (.143)	.60	1.103 (.121)	.59
Few	true	1.286 (.130)	.40	1.168 (.181)	.41
Fewer than half	false	1.114 (.104)	.53	1.123 (.280)	.51
Fewer than half	true	1.179 (.109)	.47	1.141 (.216)	.49
Many	false	1.141 (.145)	.47	1.021 (.136)	.49
Many	true	1.082 (.115)	.53	.978 (.146)	.51
Most	false	1.078 (.112)	.54	1.015 (.206)	.53
Most	true	1.043 (.134)	.46	.961 (.183)	.47
More than half	false	1.016 (.127)	.51	.919 (.143)	.50
More than half	true	.965 (.099)	.49	.909 (.302)	.50

Model fit and model parameters

Figure 7 shows that Model 7 from Experiment 1 also has a good fit to the data from both sessions in Experiment 2. Table 3.8 shows that parameter estimates are also comparable to those found in Experiment 1.

Table 3.8: Mean parameters (*SD*) for each quantifier in both sessions.

Session	Quantifier	p_0	s	V_L	V_U	a	z	*Ter*
1	*Few*	-.28	284	-.14	.14	.24	.52	.43
		(.33)	(362)	(.04)	(.04)	(.06)	(.06)	(.11)
1	*Fewer than half*	.01	457	-.14	.14	.24	.52	.43
		(.15)	(355)	(.04)	(.04)	(.06)	(.06)	(.11)
1	*Many*	-.06	284	-.19	.19	.24	.52	.38
		(.33)	(362)	(.04)	(.04)	(.06)	(.06)	(.09)
1	*Most*	.15	284	-.19	.19	.24	.52	.38
		(.21)	(362)	(.04)	(.04)	(.06)	(.06)	(.09)
1	*More than half*	.02	457	-.19	.19	.24	.52	.38
		(.15)	(355)	(.04)	(.04)	(.06)	(.06)	(.09)
2	*Few*	-.27	196	-.16	.16	.23	.51	.43
		(.28)	(296)	(.05)	(.05)	(.05)	(.04)	(.09)
2	*Fewer than half*	.001	514	-.16	.16	.23	.51	.43
		(.11)	(352)	(.05)	(.05)	(.05)	(.04)	(.09)
2	*Many*	-.03	196	-.22	.22	.23	.53	.37
		(.31)	(296)	(.06)	(.06)	(.05)	(.04)	(.07)
2	*Most*	.13	196	-.22	.22	.23	.53	.37
		(.23)	(296)	(.06)	(.06)	(.05)	(.04)	(.07)
2	*More than half*	.03	514	-.22	.22	.23	.53	.37
		(0.09)	(352)	(.06)	(.06)	(.05)	(.04)	(.07)



(a) (b)

Figure 3.7: Defective cumulative density plots show the average fit of Model 7 to data from 3.7a Session 1 and 3.7b Session 2. For this visualization, we plot the mean 5%, 15%, 25%, 35%, 45%, 55%, 65%, 75%, 85%, and 95% percentiles over participants, scaled by the proportion of true and false responses, separately for the data and the Model 7 prediction.

Stability of thresholds

The LT hypothesis predicts that thresholds should be stable over time. The CV hypothesis allows variation in thresholds for *most*, if *most* is vague in a criterial sense. To test these hypotheses, we performed Experiment 2 in which participants repeated the same experiment in two sessions two weeks apart.

To compare each parameter between the sessions, we used a Bayesian paired t test (R function *ttestBF* from *BayesFactor* library, Morey, 2018). We chose to use a Bayesian statistic because we wanted to quantify evidence for the null hypothesis. The Bayesian t test indicates the relative likelihood of a difference in parameter estimates between sessions, expressed by the Bayes Factor. A large Bayes Factor suggests that there is a systematic parameter difference between the sessions, whereas a Bayes Factor less than 1 provides evidence for the absence of a difference, suggesting stable parameters across sessions. We mostly focused on the stability of the threshold parameters (Table 3.9). We predicted that the threshold for *more than half* and *fewer than half* should be stable over time at 50%. We considered that thresholds for other quantifiers might differ between sessions.

Table 3.9: Bayes Factors for paired t tests for parameter estimates between sessions. The Bayes Factors were the same if the parameters were constrained across quantifiers: s was the same for *more/ fewer than half* and the same for *most*, *many*, *few*; V_L and V_U were symmetric (V_L = -V_U); a was the same for all quantifiers; z and *Ter* were the same for positive and the same for negative quantifiers.

Quantifier	p_0	s	V_L	V_U	a	z	*Ter*
Few	.14	.41	71	71	.63	.39	.15
Fewer than half	.16	.19	71	71	.63	.39	.15
Many	.17	.41	1106	1106	.63	.15	.16
Most	.22	.41	1106	1106	.63	.15	.16
More than half	.17	.19	1106	1106	.63	.15	.16

We found that the Bayes Factor for *most* was 0.22 and for *more than half* 0.17, which indicates evidence in favor of the null hypothesis (no difference between parameters between sessions). In particular, the Bayes Factor for individual threshold parameters was below 0.33 for all quantifiers, which indicates substantial evidence in favor of the null hypothesis (Harold Jeffreys, 1961). These results speak in favor of threshold stability.

We also tested the stability of other DDM parameters between two experimental sessions (Table 3.9). In general, the model parameters were stable across sessions, apart from V_L and V_U parameters. We tested whether participants accumulated evidence faster in the second session, by computing the maximum speed of evidence accumulation in both sessions, which was operationalized as the distance between asymptotes (V_U – V_L) and tested the difference in distance

between sessions (Bayesian paired t test, R function *ttestBF* from *BayesFactor* library, Morey, 2018). We found substantial evidence in favor of the hypothesis that participants speed up the evidence accumulation process in the second session for positive quantifiers (BF = 1106 ± 0%) and for negative quantifiers (BF = 71 ± 0%). This difference can be explained in terms of the training effect, consistent with previous literature that found that training effects are reflected in increased drift rates (Dutilh, Krypotos, & Wagenmakers, 2011; Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009; Petrov, van Horn, & Ratcliff, 2011).

Correlation of parameters between sessions

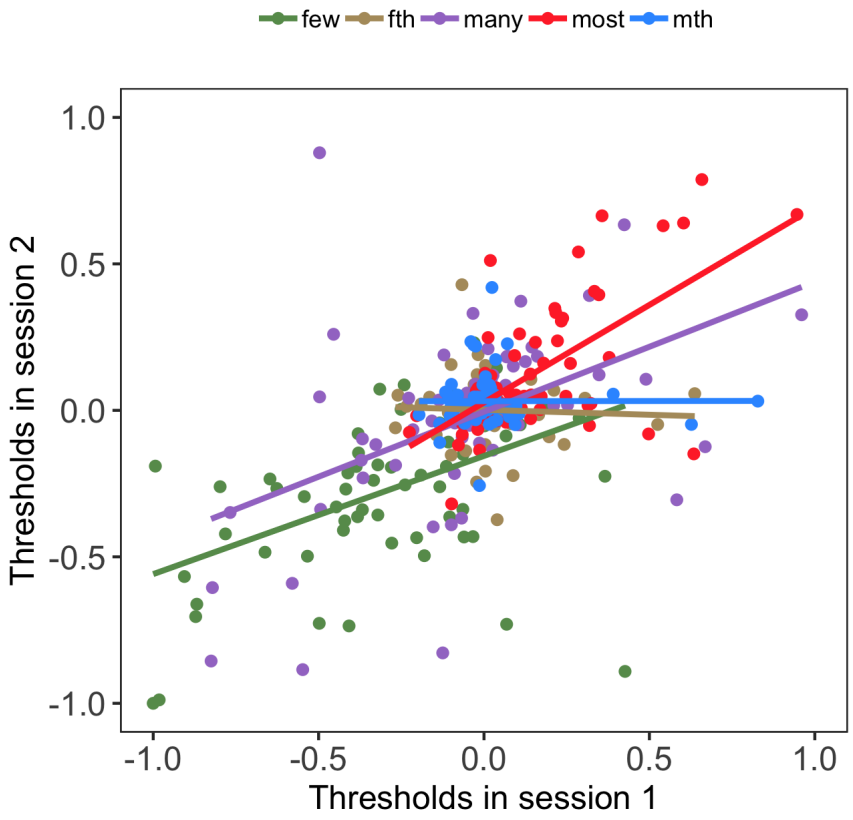


Figure 3.8: Scatterplots from Experiment 2 showing correlations of thresholds between Session 1 and 2: *few* ($r(62) = 0.48$; $p < 0.001$), *fewer than half* (fth) ($r(62) = -0.05$; $p = 0.72$), *many* ($r(62) = 0.47$; $p < 0.001$), *most* ($r(62) = 0.63$; $p < 0.001$), and *more than half* (mth) ($r(62) = 0.002$; $p = 0.99$).

In addition, we correlated the thresholds parameters across sessions (Table 3.10 summarizes all correlations). The correlations between thresholds for vague quan-

tifiers (*most*, *many*, and *few*) were moderate and significant (Figure 3.8). The correlations between thresholds for quantifiers with sharp meaning boundaries (*more than half* and *fewer than half*) were close to zero (Figure 3.8). Although this pattern of results seems counterintuitive, the very low correlations for *more than half* and *fewer than half* between thresholds reflect the very low variation in thresholds for these quantifiers. The correlations for a, *Ter*, V_L, V_U, and z (for positive quantifiers) parameters are moderate or high (Table 3.10).

Table 3.10: Correlations ($df = 62$) of DDM parameters between Session 1 and 2. Note that some parameters were constrained between quantifiers (see modeling section) and therefore the correlations were the same.

Quantifier	p_0	s	V_L	V_U	a	z	*Ter*
Few	.48***	.03	.55***	.55***	.77***	.04	.53***
Fewer than half	-.05	-.17	.55***	.55***	.77***	.04	.53***
Many	.47***	.03	.51***	.51***	.77***	.31*	.46***
Most	.63***	.03	.51***	.51***	.77***	.31*	.46***
More than half	.002	-.17	.51***	.51***	.77***	.31*	.46***

*** $p < .001$; ** $p < .01$; * $p < .05$.

3.4 Discussion

In this paper, we modeled a linguistic task as a decision-making task. We chose the quantifier verification task and two widely discussed quantifiers - *most* and *more than half* - for our case study. We used the DDM to distinguish different meaning aspects such as truth-conditional representation and vagueness. Our model accounted for vagueness by assuming a noisy decision process; and individual differences in meaning representations, by estimating parameters for each participant. We tested the predictions of logical theories about truth conditions of *most* and showed that these theories need an extension to account for individual differences in representations of *most*.

The logical theories make a strong and clear prediction about the meaning representation of *most*, which should be truth-conditionally equivalent to *more than half*. The empirical data (Kotek et al., 2015; Solt, 2016) showed that *most* is used and verified differently than *more than half*. These patterns of results had multiple explanations (Carcassi & Szymanik, 2021; Hackl, 2009; Pietroski et al., 2009; Solt, 2016). None of these explanations, however, considered the difference in truth conditions between *most* and *more than half*. After controlling for pragmatic and processing effects, and accounting for vagueness and individual differences, we observe a difference in threshold between *most* and *more than half*, which is attributable to an inequivalence in truth conditions. Our results question the typical formulation of truth conditions for *most* on the basis of logical theories.

Recall the example in the introduction. If *most* has a truth condition formulated as in Example (2) then it should also have a 50% threshold, like *more than half*. Solt (2016) claimed that *most* has a truth condition as in Example (2), but it is preferably used with higher proportions. In order to be judged as *most*, the proportion of As that are B has to be "significantly" higher than a proportion of As that are not B. This happens because *most* is represented on a semi-ordered scale, on which the proportions are compared using ANS. *More than half* is represented on a ratio scale, on which proportions can be compared precisely. For the approximate comparison, the proportions have to be "significantly" greater to be distinguished from each other.

In the current study, we used an experimental paradigm, which makes this explanation unlikely. Firstly, in our experiment, participants had to compare precise proportions given as a number. The scale on which they represented proportions had to be the same for both quantifiers. Moreover, as the numbers were precise and there was no time pressure, there was no reason for participants to use ANS for *most*. Taking this into account, we endorse the conclusion that the difference between *most* and *more than half* in our experiment is due to differences in meaning representations rather than processing strategies. More specifically, we claim that for both quantifiers, participants compared the given proportion to their internal threshold, but for *more than half* this threshold was 50%, while for *most* it varied between participants. The meaning representation of *most* is, therefore, similar to the representation of the proportional *many*. When participants verify the proportional *many*, they compare the given proportion to the threshold (Partee, 1989). For both *most* and *many*, the threshold varied between participants.

In addition to testing the between-participants variability in truth conditions of *most*, we also evaluated their stability over time. Previous studies (Verheyen, White, & Égré, 2019) showed that semantic categories change over time. This is the first study that addressed the stability of semantic representations of functional words. We showed that participants' truth conditions are stable over a short period. Verheyen, White, and Égré (2019) suggested that interindividual differences relate to stable differences between groups that apply different categorization criteria. For example, the level of education (Verheyen & Storms, 2018) or individual traits (Verheyen et al., 2018) affect the choice of category criteria. Further studies are needed to explain the cause of the interindividual variation in thresholds for vague quantifiers. The intraindividual differences, in turn, speak to the probabilistic nature of the categories themselves. In the case of quantifiers, intraindividual differences would suggest that participants have access to many meanings of vague quantifiers. Our data do not support this interpretation. We showed that participants' representations of meaning were stable over time.

In both Experiment 1 and its replication (see Appendix B), we found that *most*, but not *more than half*, was verified slower when the proportions were close to 50%. Our modeling data reflected the effect of proportion on verification of

most. We discovered that *most* and *more than half* differ in growth rate parameters (see Appendix B), which we interpreted as a measure of degree vagueness. The growth rate models the steepness of the drift rate curve. For *more than half*, the drift rate had the shape of a step-like function, which indicates that the evidence accumulation process was equally fast for all proportions. For *most*, in turn, the drift rate has a smoother shape, indicating a slower evidence accumulation process around the threshold. Our data support the predicted drift rates in Figure 3.2: the blue line for *more than half* and higher threshold and greater DV for *most* (middle red line). The modeling of the growth rate shows that *most* is more akin to *many* than *more than half*.

Our exploratory analysis showed that the choice of threshold might affect reaction times for vague quantifiers (*most*, *many*, and *few*). In the case of vague quantifiers, this analysis also showed that participants were faster when the verified proportion was further from their threshold than when it was close. We should consider these findings with caution because the effect did not fully replicate (see Appendix B).

By presenting proportion as a number, we ruled out the processing explanation of differences between *most* and *more than half*. By using pseudowords, we limited the pragmatic effects. Nonetheless, one could argue that participants constructed a context for the experiment, especially because they had to verify the context-dependent quantifiers such as *many* and *few*. Although we cannot completely rule out this possibility, we argue that this is not very likely. Firstly, participants did not assign any meaning to pseudowords (indicated by a very small amount of variance explained by by-item random intercepts, see Appendix B). Secondly, the second experiment shows stable threshold parameters for all quantifiers indicating that participants probably did not construct contexts *ad hoc*.

We also note a few limitations of this study. Firstly, we did not find a single model that would fit best for all participants. In our analysis, we account for this fact by including BMA parameters. However, we cannot conclude that differences between representations of *most* and *more than half* in threshold and growth rate are the only possible sources of interindividual variations. Secondly, although we obtained a good model fit, we noticed that the model sometimes failed to predict long reaction times. The worse model fit for long reaction times is not surprising because long reaction times are rare and, therefore, difficult to predict accurately; however, in our experiment, long reaction times mostly drove the proportion effect. Thirdly, we notice that our regression models did not meet all mixed-effects model assumptions even after log-transformation of the reaction time variable. Finally, we did not replicate all the thresholds effects on reaction times. Therefore, we can only draw a limited conclusion about the relationship between degree and criteria vagueness.

The current study shows that computational modeling is necessary to understand complex linguistic behavior. Our modeling data showed apparent differences between *most* and *more than half*, but also between negative and posi-

tive quantifiers (see Appendix B). This finding indicates that we can formulate testable hypotheses about the DDM parameters to answer other linguistic questions beyond the current case study. For example, Schlotterbeck et al. (2020) linked the difference in non-decision time with an extra step in the verification of negative quantifiers, and drift rate with difficulties of processing negative quantifiers. Furthermore, the starting point can model the response bias in the different contexts. The DDM with the implemented ANS model (Ratcliff & McKoon, 2018) can be used to test how quantifiers interact with different cognitive systems (e.g., approximate and precise number systems) and how verification changes with changing task demands (Register et al., 2018). To conclude, we presented a fruitful approach to systematically study linguistic phenomenon by means of the model of the decision-making process.

Chapter 4

Representational complexity and pragmatics cause the monotonicity effect[1]

Abstract Psycholinguistic studies have repeatedly demonstrated that downward entailing (DE) quantifiers are more difficult to process than upward entailing (UE) ones. We contribute to the current debate on cognitive processes causing the monotonicity effect by testing predictions about the underlying processes derived from two competing theoretical proposals: two-step and pragmatic processing models. We model reaction times and accuracy from two verification experiments (a sentence-picture and a purely linguistic verification tasks), using the diffusion decision model (DDM). In both experiments, verification of UE quantifier *more than half* was compared to verification of DE quantifier *fewer than half*. Our analyses revealed the same pattern of results across tasks: both non-decision times and drift rates, two of the free model parameters of the DDM, were affected by the monotonicity manipulation. Thus, our modeling results support both two-step (prediction: non-decision time is affected) and pragmatic processing models (prediction: drift rate is affected).

4.1 Background and goals

Psycholinguistic studies have repeatedly demonstrated that downward entailing (DE) quantifiers are more difficult to process than upward entailing (UE) ones. While this monotonicity effect was found in a range of different cognitive tasks, such as reading and reasoning, it shows up most reliably in verification tasks (e.g.,

[1]This chapter is based on the publication: Schlotterbeck, Ramotowska, van Maanen, Szymanik (2020). Representational complexity and pragmatics cause the monotonicity effect. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3397-3403). Cognitive Science Society.

Clark, 1976; Deschamps et al., 2015; Just & Carpenter, 1971; Szymanik & Zajenkowski, 2013). Although the empirical phenomenon itself is well-documented, it is a matter of current debate which cognitive processes cause the monotonicity effect (e.g., Agmon et al., 2019; Nieuwland, 2016; Schlotterbeck, 2017). Our main aim is to contribute to this debate by testing predictions about the underlying processes derived from two competing theoretical proposals: two-step and pragmatic processing models. To this end, we model data from two verification experiments, in particular, reaction times (RT) and accuracy, using a well-established model of decision making from mathematical psychology, namely the diffusion decision-model (DDM, see e.g., Ratcliff, 1978; Ratcliff & McKoon, 2008).

4.2 Competing theoretical proposals

Various explanations of the monotonicity effect have been proposed in the literature. We distinguish between two broad classes here. Explanations in the first class (two-step processing models) are based on an additional processing step in the verification of DE vs. UE quantifiers. The earliest two-step models (e.g., Just & Carpenter, 1971) were derived from the basic hypothesis that contexts and sentence meanings are both mentally encoded in a symbolic propositional format that can then be compared to each other symbol by symbol in a verification task. The monotonicity effect is explained by the assumption of a negation symbol present in the encoding of DE but not UE quantifiers, which corresponds to an extra step in the verification process. More recent alternatives make somewhat different assumptions, e.g., about the processing of negation (cf. Kaup, Ludtke, & Zwaan, 2007) or the meaning representations involved (e.g., Deschamps et al., 2015; Schlotterbeck, 2017), but share the assumption of an additional computational step.

A radically different view is taken by accounts that rely on a pragmatic processing model (e.g., Degen & Tanenhaus, 2019), which assumes that contextual fit or pragmatic felicity is a major determinant of processing difficulty. Under this view, DE quantifiers cause processing difficulties because they are systematically dispreferred to suitable UE alternatives in various contexts (cf. Nieuwland, 2016; and also Nieuwland & Kuperberg, 2008; for an analogous view on the processing of negation) due to violation of pragmatic principles (e.g., avoidance of infrequent words or uninformative statements, cf. Grice, 1975). In order to draw an explicit connection between pragmatic considerations of this kind and data from verification tasks, verification is often thought of as production: participants in a verification task, in fact, judge whether they would utter the sentence to describe the context (e.g., Degen & Goodman, 2014; Waldon & Degen, 2020). Recent Bayesian models of rational speaker behavior (e.g., Frank & Goodman, 2012) allow us to formalize the effects of factors such as word frequencies or informativity on speakers' production probabilities. In this way, the monotonicity effect can

be explained without assuming an additional processing step (cf. Nordmeyer & Frank, 2014, for a related proposal).

4.3 Main ingredients of the DDM

In the DDM, decision processes, such as true/false judgments, are described as the accumulation of a noisy signal over time until a decision boundary is reached and a response is initiated. One main strength of the DMM is that it concurrently models both accuracies and entire RT distributions. Moreover, its free model parameters correspond to distinct components of the underlying cognitive processes. The estimation of these parameters, therefore, allows inferences about the processing components involved in the experimental task. The DDM parameters represent independent processing components, meaning that each parameter explains different RT and accuracy effects. In this way, the DDM allows to model independent sources of variation between conditions. For the present purpose, the most important parameters are *drift rate* (v) and *non-decision time* (Ter). Drift rate determines how much information is accumulated per time unit and non-decision time measures RT components that are not themselves part of the decision process, e.g., processes related to the stimulus encoding or execution of a motor response. In addition, the standard DDM model also has a parameter, a, which specifies the separation between the two decision boundaries; a parameter, z, which determines where between the two boundaries decision processes will start; and variability parameters (s_z, s_{Ter} and s_v), which allow for trial-to-trial variability of starting point, non-decision time, and drift rate, respectively. In this paper, we focus on drift rate and non-decision time parameters, which are closely related to the cognitive processes of interest. The a parameter is usually used to model speed-accuracy trade-off (fast responses, more errors vs. slow responses, fewer errors) and the z parameter to model response bias (starting points can be closer to one of the boundaries) (e.g., Mulder et al., 2014; Ratcliff & McKoon, 2008). These two parameters do not explain the typical patterns of RT and accuracy in verification of DE and UE quantifiers.

The DDM is a theoretically well-founded model (e.g., Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006) that has been applied successfully to a large variety of decision tasks (for review, see e.g., Mulder et al., 2014; Ratcliff & McKoon, 2008). For example, a good model fit was observed in previous studies that applied the DDM to RT and accuracy collected in number comparison tasks (e.g., Dehaene, 2007; Ratcliff & McKoon, 2008). As there are close similarities between number comparison and verification of proportional quantifiers, the DDM is, therefore, also a natural choice to model the latter task as. These previous studies found that drift rate is monotonically related to numerical distance, with larger drift rates for numerosities that are further apart from each other. In comparison tasks that involved the precise comparison of numerals, a step-like relationship was

observed. For approximate numerosities, drift rates were in a linear relationship with the logarithm of the ratio (log ratio) of the two numerosities involved. These findings are consistent with current theories on the representation and processing of precise and approximate numbers (e.g., Feigenson, Dehaene, & Spelke, 2004) and are also relevant for the comparison between the experiments reported below.

4.4 Link to theoretical proposals

One way to link two-step processing models to components of the DDM is to assume that monotonicity affects non-decision time in verification tasks because the truth evaluation of DE quantifiers involves an extra step in addition to the actual verification step (see Donkin, Heathcote, Brown, & Andrews, 2009, for a related discussion and empirical data from lexical decision). For example, we could think of the verification of DE quantifiers as falsification of a suitable UE counterpart followed by a subsequent, time-consuming step of truth value reversal. However, this extra step does not change the complexity of the underlying, non-negated representation and, therefore, should not affect drift rate.

By contrast, pragmatic models hold that DE quantifiers take longer to evaluate because they are generally dispreferred as descriptions of the presented contexts. Taking into account what evidence accumulation models like the DDM have revealed about processes in closely related domains, e.g., lexical selection in picture-naming tasks (e.g., Anders, Riès, van Maanen, & Alario, 2015; Anders, van Maanen, & Alario, 2019), pragmatic models let us expect that monotonicity affects drift rates: slower accumulation is expected for DE vs. UE quantifiers. This assumption is further motivated by theoretical considerations (e.g., Bitzer, Park, Blankenburg, & Kiebel, 2014; Bogacz et al., 2006) that allow us to relate parameters of the DDM (specifically drift rate) to Bayesian pragmatic models predicting utterance production probabilities from factors such as word frequencies or informativity.

4.5 Methods

We conducted two web-based experiments, in which we compared the verification of the UE quantifier *more than half* (*mth*) to the DE quantifier *fewer than half* (*fth*). We decided to use two different paradigms – one visual (i.e. sentence-picture) and one purely linguistic (i.e. sentence-sentence) verification task. By comparing these two paradigms, we were able to not only test the robustness of the effects but also their linguistic relevance. In particular, the sentence-picture experiment involves both linguistic and visual processing. By showing that similar effects occur in both setups we provide additional evidence for the linguistic character of the effects. Additionally, while the purely linguistic experiment may

rely more on the precise comparison of the numerosities involved, the visual experiment most likely relies on approximate numbers (see Szymanik, 2016, for discussion). Hence, our results also show that the monotonicity effect is not restricted to only approximate or precise processing of numerosities (cf. Dehaene, 2007). In both experiments, we collected the participants' responses and RT.

4.5.1 Participants

For the linguistic experiment, we collected data from 90 participants via Amazon Mechanical Turk (compensation: USD 4). The final sample (see "exclusion criteria") included 72 native English speakers (24 female, mean age 35; $sd = 11$; range: 22–59). Participants in the visual experiment were recruited via `prolific.co` (compensation: GBP 7.5). Data from a total of 96 native English speakers was collected, and after exclusion, the final sample consisted of 56 participants (49 female; mean age 36; $sd = 13$; range: 18–69).

4.5.2 Design, materials, and procedures

Linguistic experiment (N=72, 50 trials per quantifier):

Participants were presented with two sentences: a simple quantified sentence of the form "Q of the As are B," where "Q" was either *mth* or *fth* and "As" and "B" were pseudowords (e.g., *glerbs* and *fizzda*) generated from English nouns and adjectives (Keuleers & Brysbaert, 2010); and a sentence of the form "X% of the As are B," where "X%" was a precise percentage between 1% and 99%, excluding 50%. The original six-letter nouns and adjectives were controlled for frequency (Zipf value: 4.06; van Heuven et al., 2014). The generated pseudowords were assessed by a native English speaker. In each trial, participants saw a different pair of pseudowords. We also included filler trials with the quantifiers *most*, *many*, and *few*. For *mth* and *fth* percentages were counterbalanced between percentages above and below 50%. Participants read the first sentence self-paced and their task was to decide whether the first sentence was true given the information from the second sentence. They responded by pressing one of two response keys on their keyboard. The experiment started with a short training block consisting of 8 trials with quantifiers that were not presented in the main experiment (i.e. *some*, *all*, *none*).

Visual experiment (N=56, 240 trials per quantifier):

Participants first read a sentence such as *"more than half of the dots are blue"* self-paced and then evaluated it against a visual display showing blue and orange dots. Participants were instructed to judge as fast as possible whether the sentence was an appropriate description of the quantitative relationships depicted. They provided their response by pressing one of two keys on their keyboard. A factorial

within-participants design was used, in which the two factors MONOTONICITY (2 levels: *mth* vs. *fth*) and RATIO of the colored dots (4 levels: 28:20, 26:22, 22:26 and 20:28) were crossed, yielding eight conditions. Each participant saw 60 trials in each condition, amounting to a total of 480 trials. 480 pictures were generated by drawing colored dots at random positions in the two halves of a gray 512×256 background. The dots had a mean radius of 5.5 (drawn from a normal distribution with $sd = 1$ and were then clipped to the range $[1, 10]$). Which color was presented on which side of the picture was counterbalanced between items. Participants saw the same set of 60 pictures in the same conditions. In half of the items, the target color was blue, and in the other half, it was orange. Materials were presented in a random order and distributed across four blocks. Each block consisted of roughly 120 trials, but the precise lengths of the four blocks were chosen randomly for each participant. In between blocks, there were self-paced breaks that participants initiated by pressing a key that they did not use otherwise. We recorded which key was pressed and thereby used the breaks as 'catch trials'. At the beginning of the experiment, there was a short practice session consisting of eight trials that were similar to the experimental trials but contained different quantifiers. In total, the visual experiment took participants about 40 minutes on average, roughly twice as long as the linguistic experiment. In both experiments, participants were randomly assigned to one of two possible response key mappings.

4.5.3 Exclusion criteria

Since data were collected over the web, we applied rather strict exclusion criteria in order to ensure high quality of the final data sets. These criteria were specified in advance and were based on the specifics of the two experiments (for discussion of data exclusion in the context of web-based experiments, see Kochari, 2019). In the linguistic experiment we excluded participants if they had more than 50% responses below 300 ms (fast guesses) or did not have an increasing probability of saying 'true' ('false' for DE quantifiers) with an increasing percentage (monotonicity violation). In addition, we excluded one more participant who participated in a very similar study previously. Altogether we excluded 18 participants.

In the visual experiment, the following criteria resulted in the exclusion of 40 participants. Participants were excluded if they had extraordinarily long reading times or RT (i.e. several minutes) in some trials; if they had more than five RT above 15 s or more than five reading times above 25 s; or if accuracy was not significantly above chance in more than one condition. In addition, we checked for participants that had many fast guesses or missed more than one of three catch trials (see procedure). All of the latter had, however, already been excluded by one of the other criteria.

In the linguistic task, we also excluded trials with RT faster than 300 ms or longer than mean+2*SD (calculated for true and false responses separately). In

the visual task, we excluded trials with reading times or RT shorter than 200 ms or longer than mean+3.5*SD (calculated per participant and condition).

4.5.4 Regression analyses and modeling strategy

First, the data were analyzed using mixed-effects regression models that mainly tested for two known effects: the monotonicity effect and the interaction between monotonicity and truth value (e.g., Just & Carpenter, 1971). To this end, independent variables were recoded in the following way. The analysis of the linguistic task included the absolute value of the normalized percentage (z-scored percentage with 50% as zero) as a numerical predictor and the analysis of the visual task included the absolute value of the logarithm of the ratio of the two numerosities presented in each trial (ABSOLUTE LOG RATIO) as a factor (levels: .167 vs. .336). In addition, analyses of both tasks included the factors MONOTONICITY (levels: *fth* vs. *mth*) and TRUTH VALUE (levels: *true* vs. *false*). Conditions with *mth* were coded as *true* if normalized percentage or log ratio was positive and as *false* if they were negative. For *fth*, TRUTH VALUE was coded the opposite way.

Afterwards, the DDM was applied to test the above predictions. We fitted the DDM to data from the two experiments separately. To this end, we used the R package `rtdists` and performed maximum likelihood estimations of DDM parameters using particle swarm optimization. We estimated non-decision time (Ter), starting point (z), boundary separation (a), and drift rate (v). All variability parameters were set to 0. We assumed that log-ratio and normalized percentage are monotonically related to drift rates and specified this relationship using the following generalized logistic regression function, where: V_L is a lower asymptote; V_U is an upper asymptote; s is a growth rate; p_0 is a midpoint; and p is normalized percentage or log-ratio.

$$v(p) = V_{\mathrm{L}} + \frac{V_{\mathrm{U}} - V_{\mathrm{L}}}{1 + e^{-s(p-p_0)}} \tag{4.1}$$

4.6 Results

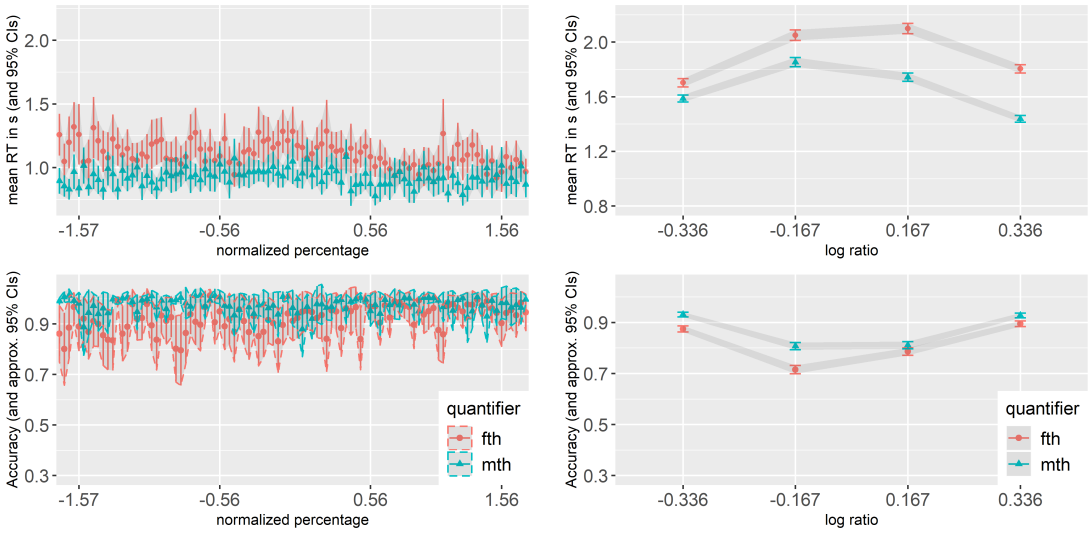
Mean RT and accuracies are shown in Figure 4.1. Below we report the results of the regression and DDM analyses[2].

4.6.1 Regression Analyses

The main results of the regression analyses are given in Table 4.1. The MONOTONICITY effect as well as the MONOTONICITY×TRUTH VALUE interaction were

[2]The data and analysis scripts of both experiments are made available at `https://osf.io/4d69v`.

replicated in RT and accuracy in both experiments. Mean RTs were faster and accuracy was higher for *mth* than for *fth* (LINGUISTIC: 926 ms vs. 1110 ms and 97.7% vs. 92.3%; VISUAL: 1655 ms vs. 1913 ms and 86.9% vs. 81.8%). Moreover, these effects were more pronounced in the false than in the true conditions (LINGUISTIC: true: 233 ms and 7.7% difference; false: 125 ms and 3% difference; VISUAL: true: 289 ms and 7.5% difference; false: 231 ms and 2.9% difference). To test for effects of MONOTONICITY independently of TRUTH VALUE, we conducted separate analyses for the true and false conditions. The MONOTONICITY effect was significant in all cases.



(a) (b)

Figure 4.1: Descriptive results. 4.1a: linguistic task; 4.1b: visual task.

Table 4.1: Results of regression analyses. MON: effect of monotonicity; INT: truth value × monotonicity interaction.

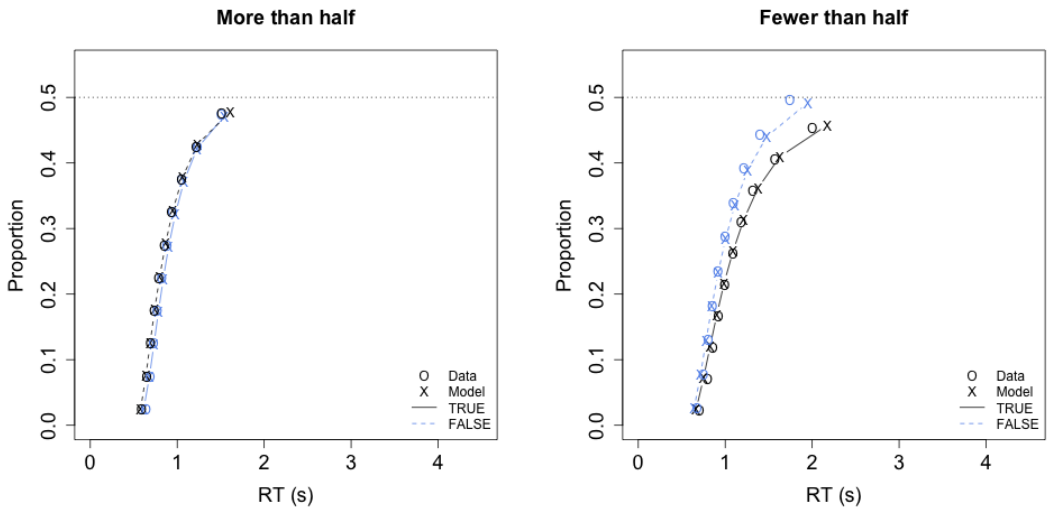
	RT						Accuracy					
	linguistic task			visual task			linguistic task			visual task		
	β	t	p	β	t	p	β	z	p	β	z	p
MON	136	5.55	$< .001$	231	15.32	$< .001$	1.01	5.09	$< .001$	.19	3.28	.001
INT	107	5.50	$< .001$	26	2.51	.012	.55	2.11	.035	.32	4.54	$< .001$
true conditions only												
MON	224	7.66	$< .001$	307	14.47	$< .001$	1.19	3.51	$< .001$	.55	9.24	$< .001$
false conditions only												
MON	151	5.71	$< .001$	246	11.44	$< .001$	1.32	3.27	.001	.14	2.24	.025

4.6.2 DDM analyses

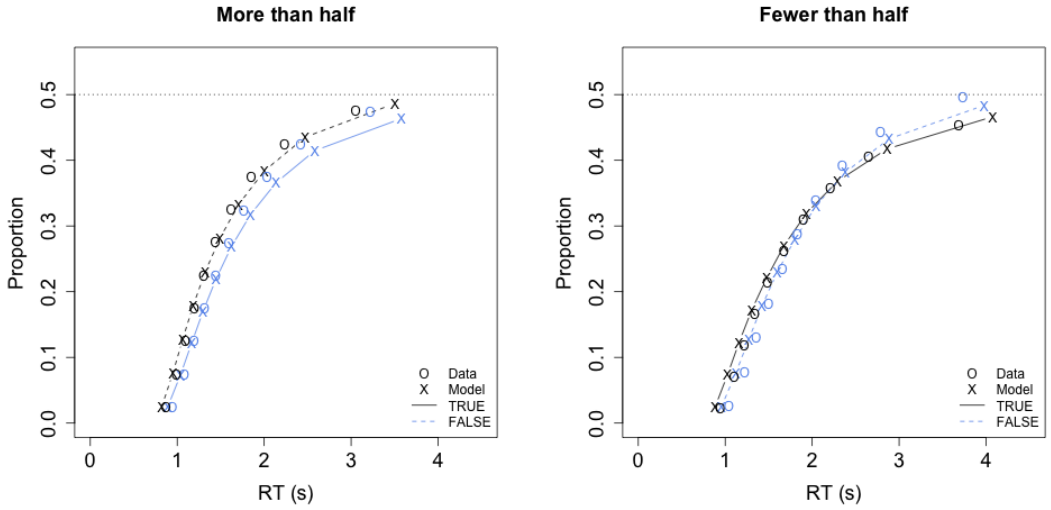
First, we fitted the DDM to the linguistic data and used model comparisons (based on BIC Schwarz, 1978) to determine which parameters differed between quantifiers (see Table 4.2). We predicted that both quantifiers should have a 50% midpoint (p_0 parameter) and growth rate (s parameter), because the truth conditions for both quantifiers were unambiguously specified. Based on the patterns of RT and accuracy for both quantifiers, we did not find evidence for a speed-accuracy trade-off, typically modelled by the a parameter (Mulder et al., 2014; Ratcliff & McKoon, 2008). Therefore, we also constrained a to be the same for both quantifiers. Additionally, we tested that the constrained parameters did not differ between quantifiers s ($t(71) = .42; p = .68$), p_0 ($t(71) = -.96; p = .34$) and a ($t(71) = -1.45; p = .15$). The final model was the best model for 66 out of 72 participants. Then, we applied the same model to the visual data. We verified that the model fit was good by examining participants individually. The overall model fit is shown in Figure 4.2.

Table 4.2: Summary of model constraining procedure

Model number	1	2	3	4
Constrained parameters	–	s	s, p_0	s, p_0, a
Number of free parameters	14	13	11	10
Model was best for:	0	1	5	66



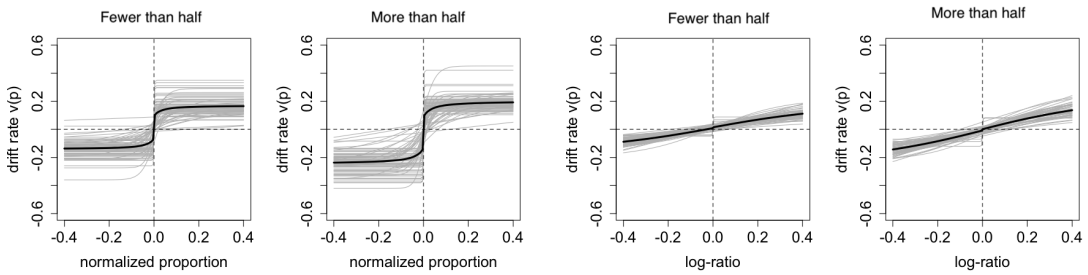
(a) (b)



(c) (d)

Figure 4.2: Defective CDF plots (Ratcliff, 1979) showing average model fit (4.2a, 4.2b: linguistic task; 4.2c, 4.2d: visual task).

In line with previous results (Dehaene, 2007), a comparison revealed that decision processes differed between the two tasks: Drift rate increased gradually with log-ratio in the visual task, whereas a step-like relation was found in the linguistic task (see Figure 4.3). Apart from this difference, we found consistent results across the two tasks. In both tasks, non-decision times were longer for *fth* than for *mth* (LINGUISTIC: $t(71) = 5.53; p < .001$; VISUAL: $t(55) = 5.74; p < .001$). The mean difference between *fth* and *mth* was 34 ms in the linguistic and 43 ms in the visual task.



(a) (b)

Figure 4.3: Relationship between drift rates and numerical information. 4.3a: linguistic task; 4.3b: visual task. Gray lines correspond to individual participants; black lines are based on mean parameter estimates.

To test for differences in drift rates, we calculated distances between the asymptotes (V_U - V_L) of the logistic regression function. We found that the

mean distances between asymptotes were larger for *mth* (LINGUISTIC: .46; VISUAL: .64) than for *fth* (LINGUISTIC: .31; VISUAL: .46). This means that drift rates were higher for *mth* than for *fth* (LINGUISTIC: $t(71) = 9.10; p < .001$; VISUAL: $t(55) = 8.46; p < .001$).

Moreover, we also tested for differences in relative starting points. In the linguistic task, we found a *yes*-bias for *mth* (the starting point was closer to the upper decision boundary) compared to *fth* (.56 vs. .49; $t(71) = 5.56; p < .001$). In the visual task, both quantifiers exhibited a *yes*-bias (.54 vs. .55; $t(55) = -.96, p = .34$).

Because Model 4 was the best model for only 66 out of 72 participants, we tested additionally if the variation between participants in best model fit had an effect on our results. To test this we computed Bayesian model averaged (BMA) parameters. The BMA method takes into account parameters from all fitted models and computes weighted average parameters according to the models' BIC values (Wagenmakers & Farrell, 2004). The BIC weight w for model i is defined by the following equation, where $\Delta_i(BIC) = BIC_i - min(BIC)$.

$$w_i(BIC) = \frac{exp\{\frac{-1}{2}\Delta_i(BIC)\}}{\sum_{k=1}^{K} exp\{\frac{-1}{2}\Delta_k(BIC)\}} \tag{4.2}$$

We tested the difference between DE and UE quantifiers in non-decision time and drift rate parameters. We found the expected difference in non-decision time ($t(71) = 5.63; p < .001$) and drift rate ($t(71) = 9.50; p < .001$). These findings indicate that the variation between participants was negligible.

4.7 Discussion

We applied the DDM to data from two web-based verification experiments in order to test predictions derived from theoretical accounts of the monotonicity effect. From two-step accounts, we derived the prediction that non-decision time would be affected, and from pragmatic processing models, we derived the prediction that drift rate would be affected.

The monotonicity effect was replicated in both experiments, and our modeling results are entirely consistent across both experiments: we found that the monotonicity manipulation affected both parameters – drift rates and non-decision times – in the expected direction. Therefore, our results support both hypotheses and indicate two potential sources of the monotonicity effect that map onto different DDM parameters. Moreover, they show that the monotonicity effect and its cognitive correlates are robust across various linguistic tasks, strongly suggesting that they are inherent in language processing. We acknowledge that an unambiguous mapping from effects in non-decision times and drift rates to representational complexity and pragmatic processes, respectively, can be challenged.

Nevertheless, our modeling results render accounts that explain effects on only one of the two parameters implausible, or at least incomplete.

Recently, Agmon et al. (2019) arrived at similar conclusions analyzing mean RT. They compared verification of quantifiers, e.g., *mth* vs. *fth*, to the verification of expressions containing positive vs. negative adjectives, e.g., *a large* vs. *a small proportion*. Like *fth*, *a small proportion* is also negative, but it is not DE. Across a range of comparable expressions, they found larger RT differences between pairs that differ along both of these dimensions, than between expressions that differ only in negativity. They argued that both negativity and downward monotonicity are sources of increased processing difficulty. One way to explain these findings in our present terms and also to explain the two sources of processing difficulty observed in estimated DDM parameters, would be to assume that negativity affects pragmatics. In contrast, only DE expressions involve an extra processing step. While the relevant theoretical distinctions are, in fact, more subtle than what we can cover here (see also Bott, Schlotterbeck, & Klein, 2019, for discussion), the empirical question of how our modeling approach relates to these findings is interesting in its own right. We plan to address this question in ongoing efforts.

Another well-documented effect – the interaction between monotonicity and truth value – was also replicated in our experiments. Classical explanations of this effect are based on verification procedures (Barwise & Cooper, 1981; Szymanik & Zajenkowski, 2013; Deschamps et al., 2015). While the observed differences in mean RT, as well as our regression analyses, are consistent with previous findings, our modeling results are unexpected under those accounts: what our results indicate is a tendency to answer "yes, true" to *mth* in the visual and linguistic task. To obtain a better understanding of how response biases are related to the interaction between monotonicity and truth value, a comparison to the processing of negation may be instructive, where a similar interaction is often observed (Just & Carpenter, 1971).

Beside the similarities mentioned, we also found differences between the two tasks. As reflected in higher RT and lower accuracy, the visual task was the more difficult among the two. Moreover, the signature of the decision processes also differed between tasks (see Figure 4.2). These findings are consistent with existing studies (Dehaene, 2007) that applied the DDM to number comparison tasks involving either approximate (dot pictures) or precise numerosities (numerals). The fact that the present analyses replicate these results indicates that our method is sensitive enough to detect qualitative differences between tasks. Thus, the consistent results on monotonicity receive indirect validation.

Finally, our results demonstrate that decision models, like the DDM, are applicable to data collected over the web. We will take a closer look at this by comparing our results to a replication in the lab.

Chapter 5

Discovering stages of processing in quantified sentences[1]

Abstract The sentences "*More than half* of the students passed the exam" and "*Fewer than half* of the students failed the exam" describe the same situation, and yet the former one is easier to process than the latter one, as reflected by shorter reaction times in the verification task. The two-step model explains this result by postulating that negative quantifiers contain hidden negation, meaning that *fewer than half* is represented roughly as *not more than half*. This account predicts an extra processing stage for negative quantifiers. To test this theory, we applied the hidden semi-Markov model multivariate pattern analysis (HsMM-MVPA) to EEG data from a picture-sentence verification task. We estimated the number of processing stages during reading and verification of quantified sentences ("*More than half*/*Fewer than half* of the dots are blue") that followed the presentation of colored dot pictures. We did not find evidence for the extra step during the verification of the sentence with *fewer than half*. Our analysis challenges the two-step model. We provide an alternative interpretation of our results in line with the expectation-based pragmatic account.

5.1 Introduction

In the 1960s, studies first showed that sentences with negation take longer to process than affirmatives (Wason, 1961). However, this effect cannot be straightforwardly attributed to negation itself. This is because explicit negation lengthens the sentence: the longer the sentence, the more complex it is and therefore the

[1]This chapter is based on the manuscript: Ramotowska, Archambeau, et al. (2022). Discovering stages of processing in quantified sentences (unpublished manuscript). The analyses presented in this chapter were conducted on data collected and published previously: Augurzky, Schlotterbeck, Ulrich (2020). Most (but not all) quantifiers are interpreted immediately in visual context. *Language, Cognition and Neuroscience, 35 (9)*, 1203–1222.

longer it takes to process (see Grodzinsky et al., 2020, for methodological discussion). To avoid this confound, Just and Carpenter (1971) tested three types of negation: explicit syntactic negatives (e.g., *none*), implicit syntactic negatives (e.g., *few*), and semantic negatives (e.g., *a minority*). They found that participants verified all types of negatives longer than affirmatives. Because the sentences with implicit syntactic and semantic negatives and affirmatives were of the same length, this study confirmed that the processing difficulties related to negation are not just a function of the length of the sentence, but are inherent to negation. We will refer to this effect as the polarity effect, a general linguistic phenomenon of negative expressions (including sentential negation) being more difficult to process than their affirmative counterparts (Just & Carpenter, 1971, see Clark, 1976 for review).

Several theoretical proposals aimed to explain this highly replicable effect. In this paper, we discuss and test one of the general approaches, namely the two-step model (see Clark, 1976, for review). We refer to the two-step model as a class of models that share a common assumption: they postulate an extra processing step in the verification of negation and negative expressions. To control for the confound caused by the explicit negation (the length of the sentence), we investigated the polarity effect by comparing a pair of quantifiers: positive (*more than half*) and negative (*fewer than half*).

The two-step model was inspired by studies on sentential negation (Clark, 1976; Clark & Chase, 1972; Kaup, Lüdtke, & Zwaan, 2006). It is also well-grounded in the semantic analysis of negation and negative expressions (Grodzinsky et al., 2018). It appeals to the idea that a sentence is processed in a sequence of stages. These processing stages correspond to the mental operations of building the representation of the sentence. The more complex the sentence, the more operations it involves. The main assumption of the two-step model is that negative expressions (e.g., the quantifier *fewer than half*) contain so-called hidden (or implicit) negation, which corresponds to an additional mental operation. The extra processing step contributes to the latency of sentence processing. It should therefore be reflected in the measure of the reaction times, namely, it should take longer to process negatives than affirmatives. This prediction bore out in behavioral studies on explicit negation, expressed in English by *no*, *not*, *it is not true that* (Just & Carpenter, 1971); and implicit negation, expressed by negative quantifiers (e.g., *few, fewer than half*, Schlotterbeck et al., 2020), adjectives (e.g., *short*, Tucker, Tomaszewicz, & Wellwood, 2018), or location words (e.g., *below*, Clark & Chase, 1972).

The idea that the upcoming information (for example, a sentence) is processed in a series of cognitive stages has a long tradition not only in linguistics but also in experimental psychology. As early as the late 1960s, Donders (1969) laid the foundation for analysis of processing stages. As an extension, Sternberg (1969) proposed the additive factor method to study the stages of processing in reaction time data. The processing stages postulated in cognitive models are also

reflected in the stages of processing in the brain (Zylberberg, Dehaene, Roelfsema, & Sigman, 2011).

Thus far, the two-step model has never been tested directly. The experimental studies testing the two-step model's prediction used measures of mean reaction times (e.g., Clark & Chase, 1972; Just & Carpenter, 1971; Kaup et al., 2006) or event-related potentials (ERPs) (e.g., Fischler, Bloom, Childers, Roucos, & Perry, 1983; Farshchi, Andersson, van de Weijer, & Paradis, 2020) that are not suitable for determining the stages of processing postulated by the model. Thus, these methods only indirectly showed support for the two-step model, by assuming that the reaction times or mean electroencephalographic (EEG) data patterns are due to the extra processing step. The recent advancements in computational modeling allow us to directly estimate the number of processing stages in simple cognitive tasks (Anderson et al., 2016). By using the computational model we can also estimate the number of processing stages for quantified sentences and directly test the two-step model predictions.

In this paper, we challenge the two-step explanation of the polarity effect in quantifiers by directly estimating and comparing the number of processing stages in the picture-sentence verification task with two quantifiers *more than half* and *fewer than half*. We applied the hidden semi-Markov model multivariate pattern analysis (HsMM-MVPA Anderson et al., 2016) to the EEG data to discover the stages of processing in the quantified sentences. In the next sections, we will explain the key concepts of the two-step model and present experimental findings that indirectly support its predictions. Then we will point out the limitations of these studies and show how we can directly test the two-step model by using the HsMM-MVPA method.

5.1.1 The two-step models

The 'true' and 'conversion' models of negation

Clark and Chase (1972) formulated the first model of negation processing (see also Clark, 1976), called the 'true' model of negation[2]. The model described four stages of processing in the sentence-picture verification task. In this task, participants read either affirmative or negative sentence and then they see a picture which either corresponds to the sentence or not. They have to verify the sentence based on the picture. According to the model, participants first build the representation of sentence (Stage 1); then, the representation of picture (Stage 2); next, they compare these representations (Stage 3); and finally, they respond (Stage 4). The model explained the observed reaction time differences between affirmative and negative sentences via hidden parameters, called latency components. It included a parameter related to the longer encoding time of negative (e.g., *below*) vs. affirmative (e.g., *above*) expressions in Stage 1. It

[2]The 'true' model is also known in a literature as the schema-plus-tag model.

specifies the latency components in Stage 3 for representing negation (Negation Time) and the falsification process (Falsification Time). Moreover, it also defines a procedure for comparison of the picture and the sentence representations in Stage 3. This procedure consists of a few mental operations, which correspond to processing steps.

To better understand the 'true' model, consider the following example. Participants verified the sentences "A *is/isn't above/below* B" or "B *is/isn't above/below* A" against the picture where A is placed above B, represented as (A above B). The True Affirmative sentence "A is above B" is simply represented as (A above B). The verification of this sentence takes time t_0. The sentence "B is below A," in turn, is represented as (B below A) and requires additional time to encode the negative expression *below* (t_0 + a). The False Affirmative sentence "B is above A" is represented as (B above A), and takes longer to process than True Affirmative because of the Falsification Time latency component (t_0 + c). The True Negative sentence "B isn't above A" is represented as (false(B above A)), and the representation of this sentence requires additional Negation Time to encode the negation (*isn't*), and the Falsification Time (t_0 + c + b). Finally, the False Negative sentence "A isn't above B" is represented as (false(A above B)), and requires only the Negation Time, but not the Falsification Time (t_0 + b). To summarize, the model assumes that the negative sentences are more complex to represent than affirmative sentences and predicts the interaction between affirmative and negative sentences and the truth value (predicted already by Barwise & Cooper, 1981, see also Szymanik & Zajenkowski, 2013).

The interaction between positive and negative sentences and the truth value is a crucial prediction of the 'true' model of negation (Clark, 1976) that distinguishes this model from alternative proposal, the 'conversion' model of negation (Young & Chase, 1971)[3]. The 'conversion' model assumes that the negative sentence "B isn't above A" can be converted into the affirmative sentence "B is below A" and verified after the conversion. This model postulates the Conversion Time (k) for all negative sentences (see Clark, 1976, for a detailed description of this model). When directly instructed, participants can use conversion in an effective way (Young & Chase, 1971). Under conversion, the reaction times pattern changes. While the 'true' model predicts interaction between affirmative and negative sentences and the truth value, the 'conversion' model predicts only the main effects of truth value (Falsification Time, c) and negation (Conversion Time, k), but no interaction. While some studies support the 'conversion' model (Wason, 1961), the model's application is limited to the tasks with two contradictory predicates (e.g., *odd number* vs. *even number*) where one is a negation of another (e.g., *odd number* means *not even number*).

The 'true' model brought under discussion two important concepts for the current study. Firstly, it assumes an additional processing step that could be

[3]The 'conversion' model is also known in a literature as the fusion model.

mapped on the cognitive stages. Secondly, it postulates two sources of processing difficulties. The first one is at the mental representation level, where negatives (e.g., *below*, *fewer than half*) contain the hidden negation and therefore have a more complex representation. The second source lays in the verification procedure that consists of more steps for negative sentences, reflected, in turn, in longer reaction times. These two features of the two-step model were developed in two more recent proposals (Kaup et al., 2006; Grodzinsky et al., 2018). The two-step simulation hypothesis (Kaup et al., 2006) explains the two-step process of building the representation of the negative sentence. The representational and verification complexity hypothesis (Grodzinsky et al., 2018), in turn, directly refers to two sources of processing difficulties of negative expression.

The two-step simulation hypothesis

According to the two-step simulation hypothesis, the representation of the negative sentences like "A is not above B" contains the positive proposition "A is above B", called the to-be-negated state (Kaup et al., 2006). To access the representation of the actual state of affairs, firstly, participants have to represent the to-be-negated sentence and mentally tag it as false. For example, to represent the sentence "The glass is not empty", they have to represent the sentence "The glass is empty." The simulation account explicitly postulates an extra step in the processing of the negated sentence.

The question arises when participants switch to the correct representation. To investigate this question, Kaup et al. (2006) presented participants with negative and affirmative sentences, and pictures that either matched or mismatched the actual state of affairs expressed by the sentences. The pictures were presented with a delay of 750 ms and 1500 ms. For negative sentences, when the delay was 1500 ms, participants responded faster to matching pictures than the mismatching pictures. No such effect was observed for affirmative sentences. However, when the delay was 750 ms, the facilitation effect was reversed. This finding supported the two-step simulation hypothesis (Kaup et al., 2006, 2007). With a longer delay of picture presentation, participants had more time to shift their attention from the representation of the to-be-negated state and focus on the actual state of affairs.

Two-step models and representational complexity

Grodzinsky et al. (2018) proposed two sources for the difficulties of processing negative quantifiers: the representational and verification complexity. Verification complexity is related to the interaction between polarity (positive vs. negative quantifiers) and truth value of the sentence (Barwise & Cooper, 1981; Just & Carpenter, 1971; Szymanik & Zajenkowski, 2013). Representational complexity is related to the number of downward entitling operators. For example, the

comparative *more* is represented as *many* + *er*, while *fewer* is represented as *little* + *many* + *er*, where *little* is an extra downward entitling operator not present in *more*.[4]

Grodzinsky et al. (2018) decomposed comparative *more* to *many* and a downward entailing operator *-er*, and *less* to *many* and two downward entailing operators *-er* and *little* .

The proposla by Grodzinsky et al. (2018) is another reincarnation of the two-step model. Like its predecessors, it also claims that negation or negative expressions take longer to represent because of the complexity of their representation. In line with other theories discussed, it postulates that the verification procedure is another source of longer reaction times. All two-step models make two predictions: (1) negative expressions are more costly to represent because of the complexity of the representation, and (2) verification of negative expressions should interact with the truth value, namely, the verification stages should reflect the complexity of the verification procedure.

5.1.2 Electroencephalography evidence for two-step model

Besides the evidence from reaction time experiments, the two-step model is also supported by electroencephalography (EEG) findings. The classical EEG studies on language processing use the event-related potential (ERP) technique, which involves averaging the signal over trials and participants. Two components are particulary interesting for language processing — N400 and P600. The N400 component is sensitive to semantic mismatch and incongruency (Kutas & Hillyard, 1980), as well as to world knowledge, discourse, cloze probability, and non-linguistic meaning processing (see Kutas & Federmeier, 2011, for review). It is a signature of the lexical retrieval processes (Delogu, Brouwer, & Crocker, 2019). P600, in turn, was first linked to syntactic processing (Hagoort, Brown, & Groothusen, 1993), but is also related to semantic integration (Brouwer, Fitz, & Hoeks, 2012; Brouwer, Crocker, Venhuizen, & Hoeks, 2017).

The first EEG evidence for the two-step processing of negation comes from the phenomenon called negation-blind N400 (Fischler et al., 1983). A sentence like "A dog is a fish" is false and semantically incongruent. It should, therefore, elicit the N400 potential on the final word of the sentence (*fish*). Fischler et al. (1983) showed that the N400 was induced not only by false sentences like "A dog is a fish," but also by true negative sentences like "A dog is *not* a fish," which is a correct and semantically congruent sentence. The lack of N400 reduction

[4]Downward and upward entailment refer to the entailment pattern. For sets A and A' if A⊆A' than quantifier Q is upward entailing if Q(A)⊆Q(A') and downward entailing if Q(A')⊆Q(A). For example, the sentence "*More than half* of men run fast" entails that "*More than half* of men run," while the sentence "*Fewer than half* of men run" entails that "*Fewer than half* of men run fast." Note that both operators *-er* and and *little* are downward entailing operators: *-er* is a comparative downward entailing operator, and *little* is a negation operator.

in the presence of negation was interpreted as evidence for a delay in processing. Palaz, Rhodes, and Hestvik (2020) showed a similar result in a more pragmatically felicitous context.

In another study, Dudschig and Kaup (2018) used the lateralized readiness potential (LRP) and showed that the to-be-negated information is initially activated. They argued that the clash between negated information and the actual state of the world is processed similarly to conflict in conflict-monitoring tasks (Botvinick, Carter, Braver, Barch, & Cohen, 2001; van Maanen & van Rijn, 2010; van Maanen, van Rijn, & Taatgen, 2012). The idea that negation requires switching between two mental representations was further supported by the EEG signatures of response inhibition in negation processing (Beltrán, Morera, García-Marco, & De Vega, 2019). These findings support the idea that the explicit negation is represented in two steps and that additional cognitive resources are needed to choose between the representations.

A few studies (Augurzky et al., 2020; Urbach & Kutas, 2010; Urbach, DeLong, & Kutas, 2015) tested the online processing of negative and positive quantifiers. For example, Urbach and Kutas (2010) manipulated the lexical-semantic associations between quantifiers (*most*, *few*), adverbs (*often*, *rarely*) and nouns to create typical and atypical sentences. They expected to find cross-over interaction between quantifier/adverb and typicality, as reflected by N400. What they found, however, was an asymmetry in N400 amplitude for positive vs. negative quantifiers. The N400 effect followed the predicted patterns only for positive expressions. Moreover, they found that the prefrontal positivity in atypical sentences was more pronounced for negative expressions, suggesting the negative expressions require additional processing compared to positive. In the follow-up experiment, Urbach et al. (2015) demonstrated that in the pragmatically appropriate discourse context N400 follows the expected full cross-over interaction pattern. Together, some studies (Nieuwland & Kuperberg, 2008; Urbach et al., 2015) demonstrated that negative expressions can be processed easier in a pragmatically felicitous context, while, others (Orenes, Moxey, Scheepers, & Santamaría, 2016) showed that sentences with negation are still processed slower than affirmative sentences.

Further evidence for a delay in processing of negative quantifiers comes from a picture-sentence verification task (Augurzky et al., 2020). Because previous studies (Urbach & Kutas, 2010; Nieuwland & Kuperberg, 2008) showed that discourse information can affect processing of negative quantifiers or negation, Augurzky et al. (2020) presented participants with picture-context instead of sentence-context. Pictures, in contrast to world-knowledge based sentences, were equally informative for all quantifiers in the experiment.

Augurzky et al. (2020) tested the online verification of sentences such as "*More than half* of the dots are blue," or "*Fewer than half* of the dots are yellow." They chose quantifiers well-controlled for semantic properties, namely *more than half* and *fewer than half*, instead of *few* and *most* tested in previous experiments (Urbach & Kutas, 2010; Urbach et al., 2015). In the experiment by Augurzky et

al. (2020), participants were presented with a picture and then a sentence word by word for 500 ms. The researchers tested the N400 effect in the time window 300 to 400 ms after the adjective onset. They found a contrast in the N400 for false vs. true sentences when the quantifier was *more than half* and a lack of effect for *fewer than half*. Moreover, in an exploratory analysis, they found a greater late positivity activation for *fewer than half* than for *more than half* in the time window 450 to 800 ms after the quantifier onset.

The authors proposed two interpretations of this finding. According to the first one, processing of *fewer than half* is more cognitively costly than *more than half* and the late positivity reflects an increase in attentional demands. According to the second interpretation, the positive component is related to the revision of the context update. The participants encoded the picture in terms of the larger proportion, and as soon as they saw the *fewer than half* quantifier, they had to revise the discourse model.

The analysis of the late positivity in the time window after the quantifier onset by Augurzky et al. (2020) was exploratory and did not directly show that the difficulties in processing negative quantifiers were associated with an extra processing step. However, their interpretations of late positivity could be framed in the two-step model. For example, increasing attentional demands might reflect the processing of hidden negation. The 'true' model (Clark & Chase, 1972) and representational complexity hypothesis (Grodzinsky et al., 2018) predict an extra cost of representation of negatives. In the 'true' model of negation, the encoding cost is reflected by the parameter in the sentence representation stage (Clark, 1976; Clark & Chase, 1972). Moreover, the two-step model predicts that participants prefer to encode the picture with a positive quantifier (Clark & Chase, 1972; Clark, 1976).

The two-step model found substantial support in experimental data on processing negative sentence (Clark & Chase, 1972; Clark, 1976; Just & Carpenter, 1971; Kaup et al., 2006). However, its predictions were never tested directly. In this study, we showed that the model could be tested directly, if we could estimate the number of processing stages for each experimental condition and map them on the predicted by the model number of stages. To directly test the two-step model, we used the Hidden semi-Markov Model Multivariate Pattern Analysis method developed by Anderson et al. (2016). In the next section, we elaborate on the main theoretical assumptions of this method and its relation to the traditional ERP analysis.

5.1.3 Hidden semi-Markov Model Multivariate Pattern Analysis (HsMM-MVPA)

We used the HsMM-MVPA to estimate the stages of processing in quantified sentences. Borst and Anderson (2015) proposed a new method to analyze the

EEG data that makes it possible to discover stages of processing. In contrast to traditional ERPs, the HsMM-MVPA analyzes the EEG signal at the single-trial level instead of averaging it from multiple trials and participants. The HsMM-MVPA algorithm detects so-called bumps, the negative or positive deflections in EEG signal across the scalp. The bump signals the onset of a new cognitive process, and it is followed by the flat, where the signal is characterized by zero mean amplitude. Together, bump and flat assemble a processing stage.

By identifying bumps we can identify an increase in brain activity and thus associated cognitive stages. Anderson et al. (2016) assumed that the bumps have a duration of 50 ms. Although, this assumption was somewhat arbitrary, it gave reliable results even when the bumps had slightly different durations (see appendix in Anderson et al., 2016, for discussion and mathematical details). Moreover, the HsMM-MVPA model assumes that bumps cannot overlap.

Anderson et al. (2016) used a semi-Markov model to allow for variability in stage durations. Flats can have different durations under the assumption that cognitive processes have various lengths depending on the task condition. Moreover, flats are variable from trial to trial under the assumption that information processing by participants is also prone to trial-by-trial variability. In the HsMM-MVPA, the duration of the flats is modelled as a distribution. Because the first bump might not occur exactly with the onset of the trial, the first stage starts with the flat. Therefore, for n bumps there are always $n+1$ stages.

The main advantage of the HsMM-MVPA method is that it makes it possible to infer cognitive stages directly from EEG data using all participants and trials. Depending on the complexity of the cognitive task and participants' capacities, cognitive models postulate the differences in stages duration between experimental conditions. By using the HsMM-MVPA method, we can directly test these predictions.

The potential of the HsMM-MVPA method has been shown in a number of simple cognitive tasks (see Borst & Anderson, 2021, for review). The first study that applied the HsMM-MVPA method to EEG data (Anderson et al., 2016) tested the ACT-R model predictions regarding the stages of processing in the associative recognition task and the Sternberg Working Memory task. The HsMM-MVPA was also applied to the perceptual speed-accuracy trade-off task (van Maanen et al., 2021), perceptual decision-making task (Berberyan, van Maanen, van Rijn, & Borst, 2020), and working memory task (Zhang, van Vugt, Borst, & Anderson, 2018). In the domains closer to quantification, Zhang, Walsh, and Anderson (2018) validated the HsMM-MVPA method in a mathematical problem-solving task, and Berberyan, van Rijn, and Borst (2021) discovered the stages of processing in a lexical decision task. Together, the validity of the HsMM-MVPA method is well-established in simple cognitive tasks. The methodological advancement of the current study is to apply this method to a more complex task in which participants have to process a stream of stimuli, such as the words of a sentence.

The relationship between HsMM-MVPA and ERPs

Thus far, we have outlined the HsMM-MVPA method at a conceptual level. To apply the method to the EEG data, Anderson et al. (2016) proposed a linking assumption between the EEG signal and bumps estimated by the HsMM-MVPA. The main challenge to estimate the stages of processing from the EEG signal is to solve the problem of trial-by-trial variability in the endogenous ERP components. Anderson et al. (2016) postulated that the HsMM-MVPA method identifies bumps of EEG activity, which correspond to the ERPs. This assumption is compatible with two theories of ERP generation (Makeig et al., 2002): the classical theory and the synchronized oscillation theory.

The classical theory (Shah et al., 2004) of ERPs claims that certain brain regions generate the phasic burst of activity in response to the cognitive event. The activity burst is reflected in the EEG signal as a sinusoidal peak, uncorrelated with the rest of the signal. The peak becomes visible after averaging many trials in the ERP analysis. However, the property of the peak depends on the trial-by-trial variability. The peak might disappear during averaging if the variability is sufficiently large. The synchronized oscillation theory (Basar, 1980; Makeig et al., 2002), in turn, postulates that the cognitive event causes a phase reset in a certain frequency, instead of a single peak. Nonetheless, the rest frequency can be mapped onto the sinusoidal peak (Anderson et al., 2016). Both ERP generation theories give similar or even indistinguishable patterns, as shown by simulation studies (Yeung, Bogacz, Holroyd, & Cohen, 2004; Yeung, Bogacz, Holroyd, Nieuwenhuis, & Cohen, 2007). The HsMM-MVPA is based on the classical theory for conceptual simplicity (Anderson et al., 2016). It detects bumps in the EEG signal, which signal the onsets of ERP components associated with new cognitive processes.

The limitations of ERPs

To test the two-step model predictions, we applied the HsMM-MVPA instead of classical ERP analysis. Because EEG has an excellent time resolution, it makes it possible to study how the sentence is processed incrementally (e.g., Urbach & Kutas, 2010; Urbach et al., 2015; Augurzky, Bott, Sternefeld, & Ulrich, 2017; Augurzky et al., 2020). Moreover, as we mentioned in the previous sections, some ERPs are a well-established signature of linguistic-specific processes (e.g., N400). Although these two reasons make the ERPs method particularly attractive for the linguistic research, the HsMM-MVPA method is more suitable to test the two-step model, because it can detect onsets of processing stages on a trial-by-trial basis. Two shortcomings of the ERP method limit the two-step model hypothesis testing.

Firstly, components such as N400 and P600 do not have a single interpretation. For example, according to the access/retrieval account, the N400 is related to

word retrieval and modulated by the context (e.g., Brouwer et al., 2012; Kutas & Federmeier, 2011; Delogu et al., 2019), while the integration account claims that it is a signature of integration difficulties of a word into a sentence (e.g., Hagoort, Hald, Bastiaansen, & Petersson, 2004). Moreover, some studies showed that N400 is sensitive to semantic truth value of a sentence (Augurzky et al., 2017, 2020), while others (Wiswede, Koranyi, Müller, Langner, & Rothermund, 2013) showed that the late negativity is a signature of truth value evaluation. In addition, the N400 amplitude depends on the discourse context (Urbach et al., 2015).

Secondly, the ERP analysis is always constrained to the specific time windows. The studies using the ERP method chose a time window in a somewhat arbitrary manner because the onset of a component peak can vary from one trial to another. For example, as Berberyan et al. (2021) noticed, the N400 potential has a maximum amplitude between 200 and 600 ms. However, the time window chosen in different studies varies significantly in terms of timing (250 to 450 ms in Barber, Otten, Kousta, & Vigliocco, 2013 vs. 300 yo 650 ms in Carreiras, Vergara, & Perea, 2007), and duration (100 ms time window in Augurzky et al., 2020 vs. 350 ms time window in Carreiras et al., 2007).

Together, the lack of the differences in N400 for *fewer than half* between conditions could be due to many reasons. The processing of *fewer than half* could have been delayed (Augurzky et al., 2020), but the difference could also not have been detected due to high trial-by-trial variability. Furthermore, the ERP analysis averages EEG signal amplitude in a fixed time window and, therefore, cannot provide evidence for *when* the onset of the cognitive process of interest occured. Moreover, it cannot show that the component peak was delayed because of another process that happened in a time window preceding the analysis (e.g., extra processing step). By using the HsMM-MVPA analysis, we overcame the aforementioned shortcomings of ERP studies.

5.2 Methods

We applied the HsMM-MVPA method to data from a picture-sentence verification task collected and analyzed by Augurzky et al. (2020). For a detailed description of the experimental design, procedure and EEG recording, see Augurzky et al. (2020).

We reanalyzed the EEG data to test the two-step model hypothesis directly. The two-step model explained previous experimental data (see Clark, 1976, for review), regardless of the order of presentation of the picture and sentence (Clark & Chase, 1972). Therefore it can be also used to predict behaviour in a picture-sentence verification task.

5.2.1 Participants

All 33 participants were native German speakers, right-handed, not color blind, with normal or corrected to normal vision. Augurzky et al. (2020) excluded 10 participants due to muscle or voltage artifacts, or eye movements (see Augurzky et al., 2020, for details about exclusion criteria and procedure). They included data from 23 subjects in the analysis. We excluded two further participants due to artifacts, which resulted in a final sample of 21 participants.

5.2.2 Experimental design

The experiment consisted of 320 trials, 160 with short sentences and 160 with long sentences. The short sentences had the structure *More than half of the dots are blue (Mehr als die Hälfte der Punkte sind blau)* or *Fewer than half of the dots are blue (Weniger als die Hälfte der Punkte sind blau)*, while the long sentences had the structure *More than half of the dots are blue, that are inside of the semicircle (Mehr als die Hälfte der Punkte sind blau, die innerhalb des Halbkreises sind)* or *Fewer than half of the dots are blue, that are inside of the semicircle (Weniger als die Hälfte der Punkte sind blau, die innerhalb des Halbkreises sind).* The sentences were constructed in such a way that given a picture (context), preceding the sentences, one quantifier was always true and the other false. The short sentences consisted of factorial combinations of Quantifier (*more than half*, *fewer than half*) and Truth value (*true sentence*, *false sentence*), while long sentences of Quantifier (*more than half*, *fewer than half*), Truth value (*true sentence*, *false sentence*), and Preposition (*inside*, *outside*). Together, there were 80 trials per quantifier and per sentence length. For short sentences there were 40 trials per truth value, and for long sentences 20 trials per truth value and preposition.

For each combination of Quantifier x Truth value x Preposition, at least 20 pairs of context pictures were generated using Microsoft PowerPoint. Pictures contained geometrical shapes (e.g., circles, triangles, rectangles) randomly paired with a container shape (e.g., semicircle, squares). A container shape was important for long sentences, which referred to shapes inside or outside it. The objects were always shown in two different colors.

The whole experiment was divided into 10 blocks (32 trials per block). Two versions of the experiment were generated. The second version had the reversed order of blocks of the first version (e.g., the first block of the first version corresponded to the last block of the second version). Each block contained an equal number of experimental conditions and trials were presented in a pseudo-randomized order.

5.2.3 Procedure

The experiment started with a short training block. The whole session took an average two to 2.5 hours. After the electrodes were applied, participants were seated in front of a 17-inch computer screen placed in a dimly lit, soundproof cabin. They were instructed to evaluate the truth value of the sentences by pressing the F or J keys on the computer keyboard. The keys corresponded to the answers *true* or *false* and were counterbalanced across participants.

Each experimental trial started with the presentation of the context picture in the center of the screen for 1500 ms. The sentences were presented word by word for 500 ms each. In order to prevent participants from guessing whether the displayed sentence was long or short, following the adjective presentation, the comma (after long sentences) or the period (after short sentence) was presented on a separate screen. Participants did not know whether the sentence would continue until they saw the punctuation mark.

The presentation of the complete sentence was followed by display of three question marks, indicating that participants should provide a response. After participants responded, a blank screen was displayed for 500 ms. To prevent eye-movement artefacts from contaminating the experimental trial, participants were instructed to blink when they saw the three exclamation marks displayed after each trial for 1200 ms.

Participants were instructed to provide responses as soon as possible. For additional encouragement, the experiment included a timeout procedure. The initial timeout for all participants was 1200 ms. During the experiment, the timeout was adopted to participants' responses timing by using exponentially weighted moving averages (Leonhard, Fernández, Ulrich, & Miller, 2011). Participants received feedback encouraging them to respond more quickly, i.e., the word "Faster!" (*Schneller!*) displayed on the screen, if their reaction times were longer than the timeout.

5.2.4 EEG recording

The EEG signal was recorded by 32 Ag/AgCl electrodes placed across the scalp using BIOSEMI Active-Two amplifier system in the frequency of 2048Hz. In addition, 4 electrodes (2 horizontal and 2 vertical EOG), and 2 mastoid electrodes were recorded.

5.2.5 Choice of analysis time windows

Previous studies (e.g., Anderson et al., 2016; Berberyan et al., 2020) have applied the HsMM-MVPA analysis from the onset of the stimuli until the response. Given that in our experiment each word was displayed for 500 ms, it would not be possible to include whole sentences into the analysis. This would make the model

too complex and the computation intractable. Therefore, we chose a time window based on previous analysis (Augurzky et al., 2020), from the quantifier onset until 800 ms[5]. This interval includes the time window in which Augurzky et al. (2020) found late positivity for *fewer than half*. We wanted to test whether this difference in amplitude was related to the extra processing step during the representation of the quantifier. For this analysis, we included both short and long sentences because participants could not have distinguished the sentence type at this point. Additionally, by including long sentences, we analyzed more trials and increased the power of the analysis (see e.g., Wagenmakers, 2009; Lerche, Voss, & Nagler, 2017; Boehm et al., 2018, for discussion on the number of trials for computational modeling)[6]. We analyzed two conditions corresponding to quantifiers *more than half* and *fewer than half*.

The two-step model gives two predictions of when the extra step could occur. It could either occur during the comprehension of the sentences, or during the comparison between sentences and pictures (Clark, 1976). Augurzky et al. (2020) also analyzed two time windows that corresponded to the two-step model predictions. In addition to the late positivity for *fewer than half*, they also found a greater N400 for *more than half* false sentences compared to *more than half* true sentences. They did not find this effect for *fewer than half*. They interpreted this finding as evidence for the delay in processing in negative quantifiers. While the late positivity for *fewer than half* could correspond to the extra step during the comprehension, the N400 effect could reflect the delay in comparison between the sentence and picture stage for *fewer than half*.

Therefore, we included a second analysis of the time window from the adjective onset until the response to test the possible prediction that the extra processing step would be present at the end of the sentences. It could correspond to the difference between conditions in comparison of the sentence and the picture representations. In this case, the 'true' model would additionally predict the interaction between sentence polarity and truth value. This interaction should be reflected in reaction times data and stages of processing. We included only short sentences in the analysis of the later time window, and we analyzed four conditions: *more than half* true sentences, *more than half* false sentences, *fewer than half* true sentences, and *fewer than half* false sentences[7].

[5]We conducted an additional analysis of the 500 ms time window after the quantifier onset (see Appendix C). We aimed to test how many processing stages would be associated with processing of just one word.

[6]Typically, in the HsMM-MVPA studies, there are at least 100 trials per condition (Anderson et al., 2018).

[7]We also conducted additional analysis of long sentences (see Appendix C) in the time window from the adjective onset to 900 ms after onset. In this analysis we compared the truth value evaluation of the short sentences with processing of the long sentences.

5.2.6 EEG data preprocessing

The data preprocessing consisted of two stages: initial data preprocessing and artifact rejection, and specific preprocessing needed for HsMM-MVPA. For data preprocessing, we used MATLAB R2019b and R2021a (The MathWorks, Inc.), MATLAB toolbox EEGlab 2019 and 2021 (Delorme & Makeig, 2004), and preprocessing scripts adopted from Berberyan et al. (2020).

We referenced the electrodes to mastoids. We downsampled the data to 1024 Hz and applied a 0.3Hz high-pass filter and a 20Hz low-pass filter. The filters and references were the same as in Augurzky et al. (2020). In the next step, we manually cleaned the data from the artifacts, except the eye movement-related artifacts. We interpolated the signal from noisy electrodes for 8 participants. We did not interpolate more than 15% of electrodes. Following manual artifact rejection, we applied Independent Component Analysis (ICA, *runica* algorithm Delorme & Makeig, 2004; Delorme, Sejnowski, & Makeig, 2007). We removed the components related to eye movements (usually 1 or 2 components) and components related to voltage artifacts. In this way, we removed 2 components on average.

After cleaning the data, we applied preprocessing steps specific to the HsMM-MVPA analysis. We followed the steps from Berberyan et al. (2020). Downsampling of EEG data is a necessary preprocessing step for the HsMM-MVPA to make the computations tractable. We downsampled data to 100 Hz and removed the incorrect trials. We also removed trials with too short or too long reaction times based on the mean +/- 2 SD criterion. Then, we epoched the data and applied the baseline correction of 200 ms.

Bumps magnitudes and flats are not directly estimated from electrode signal. Anderson et al. (2016) performed the Principal Component Analysis (PCA) to reduce the intercorrelations of the EEG signal. They included the first 10 components that accounted for the largest variance of data (above 90%). In the final step, we also performed the PCA. PCA is also used to handle the highly correlated brain signal. We included 10 first components which accounted for 92.99% of the variance in the time window after the quantifier onset and 91.07% of the variance in the time window after the adjective onset. The data were normalized with a z-score.

5.2.7 Statistical analysis of reaction times

The main goal of our analysis was to test the prediction that sentences with *fewer than half* have at least one more stage of processing then sentences with *more than half*. We expected the verification of sentences with a negative quantifier to take longer than with a positive quantifier, because of the extra processing step. In addition, following the two-step model predictions (Clark, 1976), we expected to find the an interaction between Quantifier and Truth value.

Therefore, we tested the differences in reaction time data. In the Augurzky et al. (2020), study the reaction times for negative quantifiers were longer than for positive ones. They tested this effect only for long sentences. We expected to replicate this result in short sentences. We selected the same trials as for the HsMM-MVPA analysis and for ERP analysis. We ran a mixed-effects model with Quantifier (*more than half*, *fewer than half*), Truth value (*true*, *false*) and Quantifier Truth value interaction predictors, and tested their effects on the reaction time data (using *lmer* function from R package *lmerTest*, Kuznetsova et al., 2017). We included the by-subject random intercept and we tested the significance of the random slope for the trial (centered, model comparison with *anova* function). Because the reaction times distribution is usually skewed with a tail of long reaction times, we used log-transformation[8]. In order to interpret the main effects, we used contrast coding. We encoded *fewer than half* and *false* conditions as -0.5 and *more than half* and *true* conditions as 0.5.

5.2.8 ERP analyses

We conducted the ERP analyses to test whether we could replicate the results from Augurzky et al. (2020) study. Augurzky et al. (2020) tested the ERPs only for the long sentences. We included the short and long sentences in the analysis of the time window after the quantifier onset, and only short sentences in the analysis of time window after the adjective onset. For both analyses, we included the same trials as for HsMM-MVPA analyses. We downsampled data to 256 Hz to have the same frequency as Augurzky et al. (2020). We applied a baseline correction of 200 ms for both time windows for consistency with HsMM-MVPA. Because we wanted to replicate the finding of Augurzky et al. (2020), we defined both time windows from the stimuli onset to 800 ms after the stimuli onset.

For statistical analysis, we used a cluster-based random permutation test (Maris & Oostenveld, 2007) in Fieldtrip (Oostenveld, Fries, Maris, & Schoffelen, 2011)[9]. The cluster-based random permutation test is a non-parametric test suitable for handling the multiple comparison problem. The multiple comparison problem arises with EEG data when multiple channels and data points are included in the statistical analysis. To solve this problem, we calculated the cluster-based statistics based on the following procedure. In the first step, for every sample (pair of channel and time point) the differences between conditions was calculated and quantified by the paired t test. The paired t test was used

[8]During the data analysis, we observed that the reaction times had somewhat binomial distributions with large proportions of long reaction times that were not fully excluded with the outliers procedure. We ran a separate mixed-effects model on reaction times which were not classified as timeouts. See Appendix C for details of this analysis.

[9]In Appendix C, we replicated the Augurzky et al. (2020) analysis in the 300 to 400 ms and 450 to 800 ms time windows after stimuli (quantifier or adjective) onset using repeated measure ANOVA.

because the experimental design was within-subject. The t test calculation did not affect the false alarm rate because it was not yet decisive for the statistical significance of the difference between conditions. In the next step, samples that had a higher t value than the 0.05 thresholds were selected and clustered. The threshold at the level of 0.05 means that only the highest 5th quantile of samples was selected. In the third step, cluster statistics were calculated as a sum of t values within each cluster. The maximum cluster statistics were chosen. Finally, the Monte Carlo method was used to obtain Monte Carlo significance probability. The so-called random partition was applied, meaning that the samples were randomly assigned to experimental conditions and the cluster statistic was calculated for those newly distributed data. This procedure was repeated 1000 times. Then, the so-called Monte Carlo p value – the proportion of statistics from randomly partitioned data that exceed the initially obtained statistic from the observed data – was calculated and compared to the convectional p value at the level 0.05. The result was significant if the Monte Carlo p value was smaller than p value equals 0.05.

For both ERPs analyses, we used the paired t test for dependent samples. For the analysis in the time window after the quantifier onset, we compared two conditions: *fewer than half* and *more than half*. For the analysis in the time window after the adjective onset, we tested interactions between Quantifier and Truth value. Therefore, firstly, we computed the main effects of Quantifier and Truth value, and in the next step, we calculated the interaction effect also using the dependent t test.

ERPs after the quantifier onset

Following the findings of Augurzky et al. (2020), we selected four regions of interest (ROIs): left anterior (ROI 1: F3, F7, FC1, FC5), right anterior (ROI 2: F4, F8, FC2, FC6), left posterior (ROI 3: CP1, CP5, C3, P3), and right posterior (ROI 4: CP2, CP6, C4, P4). We expected to find a significant difference between *fewer than half* and *more than half* around 450 to 800 ms after the quantifier onset.

ERPs after the adjective onset

For the analysis in the time window after the adjective onset, we also selected four regions of interest (ROIs): left anterior (ROI 1: F3, F7, FC1, FC5), right anterior (ROI 2: F4, F8, FC2, FC6), left posterior (ROI 3: CP1, CP5, C3, P3), and right posterior (ROI 4: CP2, CP6, C4, P4) (Augurzky et al., 2020). We expected to find a significant interaction between Quantifiers and Truth value in the 300 to 400ms time window after the adjective onset. We predicted that this difference would be reflected in the N400 for the *more than half* false sentence condition compared to the *more than half* true sentence condition, and a lack of

difference for *fewer than half*.

5.2.9 HsMM-MVPA

For the HsMM-MVPA analysis, we used MATLAB R2019b and R2021a (The MathWorks, Inc.), MATLAB toolbox EEGlab 2019 and 2021 (Delorme & Makeig, 2004), and analysis scripts adopted from Berberyan et al. (2020, 2021) (available at https://osf.io/z49me/OSF).

Firstly, we applied HsMM-MVPA analysis to the data from the time window 800 ms after the quantifier onset. In this time window, the maximum number of bumps was 15[10]. We fitted the HsMM-MVPA model separately to each quantifier. The model uses the data from all participants and trials simultaneously to estimate two parameters: the bump magnitudes and flats. Based on the 10 selected PCA components, the HsMM-MVPA model estimates 10 magnitude values for every bump. The flats have a gamma-2 distribution with a shape parameter fixed to value equals 2 and a free scale parameter. The trial-by-trial variability in stage durations is captured by the scale parameter.

To obtain the maximum likelihood, HsMM-MVPA used the expectation–maxi mization (EM) algorithm. To avoid estimation of local maximum instead of maximum likelihood, we applied the same procedure described by Zhang, Walsh, and Anderson (2018) (see also Berberyan et al., 2020, 2021). Firstly, the model fitted the maximum number of bumps (n) in the time window. In the next step, the algorithm iteratively removed one bump and fitted models with bumps (*n-1*). Then all *n-1* bumps models were compared and the best model was selected. The algorithm repeated this procedure until it fitted the model with only one bump.

The log-likelihood of the model increases as the complexity of the model (number of bumps) increases. To avoid overfitting, we used the leave-one-out cross-validation following the procedure of Anderson et al. (2016). The increasing complexity of the model was only justified when the more complex model fitted better to a significantly larger number of participants. This was assessed by a computing sign test on the number of participants for whom the log-likelihood of the more complex model increased. In this way, we chose a model that generalized across the largest number of participants. The sign test was used in a number of previous HsMM-MVPA studies (e.g., Anderson et al., 2016; Berberyan et al., 2021; van Maanen et al., 2021). As a result of the leave-one-out cross-validation, we obtained the bumps magnitudes and scale parameters of the gamma-2 distribution for each participant.

[10]We noted that 800 ms divided into 50 ms should result in a maximum number of bumps 16, not 15. The 15 bump maximal model is a result of the downsampling. The shortest time window was in some trials had 79 samples, not 80. Therefore 79 samples divided into 5 samples gave a 15 bump maximal model.

5.2.10 Statistical analysis of stage durations

According to our primary hypothesis the verification of *fewer than half* should be slower than *more than half* because of the extra processing step. Nonetheless, for exploratory purposes, we also planned to compare stage durations across conditions to test whether the extra processing step is the only source of the differences between quantifiers.

Firstly, we aimed to link the stage durations with the reaction times. We expected that some stage durations might be related to the specific cognitive processes that affect the length of reaction times, while other stages could just reflect the fix processing pattern of the upcoming input (such as encoding, motor preparation). While the latter stages are not particularly meaningful for our hypothesis, the former could give us insight into differences in quantifier processing.

Using the mixed-effects regression model, we tested whether the stage durations predicted the reaction times for each experimental condition. We did not have specific predictions about each stage. Therefore we applied a backward fitting procedure. We included all stages as predictors and then excluded the insignificant predictors one by one according to their p values (we excluded the predictors with higher p values first), until only significant predictors were left in the model. We used *lmer* function from R package *lmerTest* (Kuznetsova et al., 2017) and we log-transformed all variables in the models for consistency.

Finally, we selected stages that were significant predictors of reaction times of all experimental conditions from the previous analysis and we tested the differences in their duration between conditions: *more than half* true sentences, *more than half* false sentences, *fewer than half* true sentences, and *fewer than half* false sentences. We ran mixed-effects models on each stage with predictors: Quantifier (*more than half*, *fewer than half*), Truth value (*true*, *false*) and their interaction. The predictors were contrast coded as in the reaction time analysis. We also included the by-subject random intercept and the by-subject random slope for the trial (centered) if it was significant. We used *lmer* function from R package *lmerTest* (Kuznetsova et al., 2017) to run a regression model and *anova* function for model comparison. We log-transformed the stages distributions.

5.3 Results

5.3.1 Reaction time analysis

In order to test whether *fewer than half* was verified slower than *more than half*, we ran a mixed-effects regression model. We included by-subject random slope for trial as it significantly improve model fit ($\chi^2(1) = 255.49; p < 0.001$).

We found a significant intercept ($\beta = 5.79, t = 48.50, p < 0.001$), a significant main effect of Quantifier ($\beta = -0.15, t = -5.33, p < 0.001$), and significant interaction ($\beta = -0.15, t = -2.64, p = 0.008$). The effect of Truth value was not

significant ($\beta = -0.05, t = -1.83, p = 0.07$). The verification of *fewer than half* was slower than the verification of *more than half* (see Figure 5.1). Moreover, the effect of Truth value went in opposite direction for two quantifiers: reaction times were slower for false responses in *more than half* and faster in *fewer than half*.

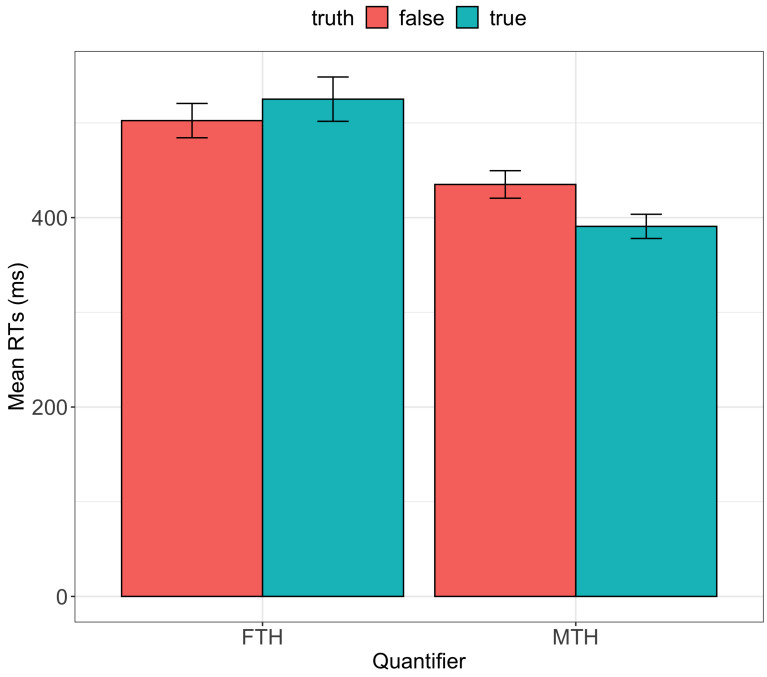


Figure 5.1: Mean reaction times for short sentences. *Fewer than half* is abbreviated as FTH and *more than half* as MTH. The error bars represent within-participant SE.

The pattern of reaction time results is compatible with the two-step model predictions (Clark, 1976). Moreover, the result justifies the processing stages analysis. The differences in reaction times should be reflected in the differences in processing stages. According to the two-step model, the reaction times for *fewer than half* are longer because of the extra processing step.

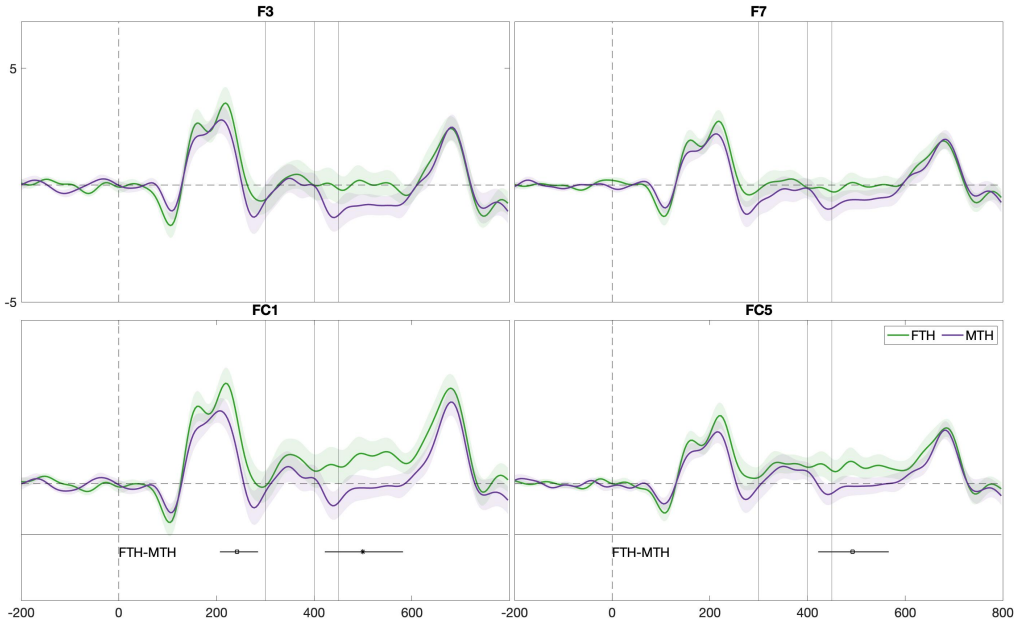
5.3.2 ERP analyses

After the quantifier onset

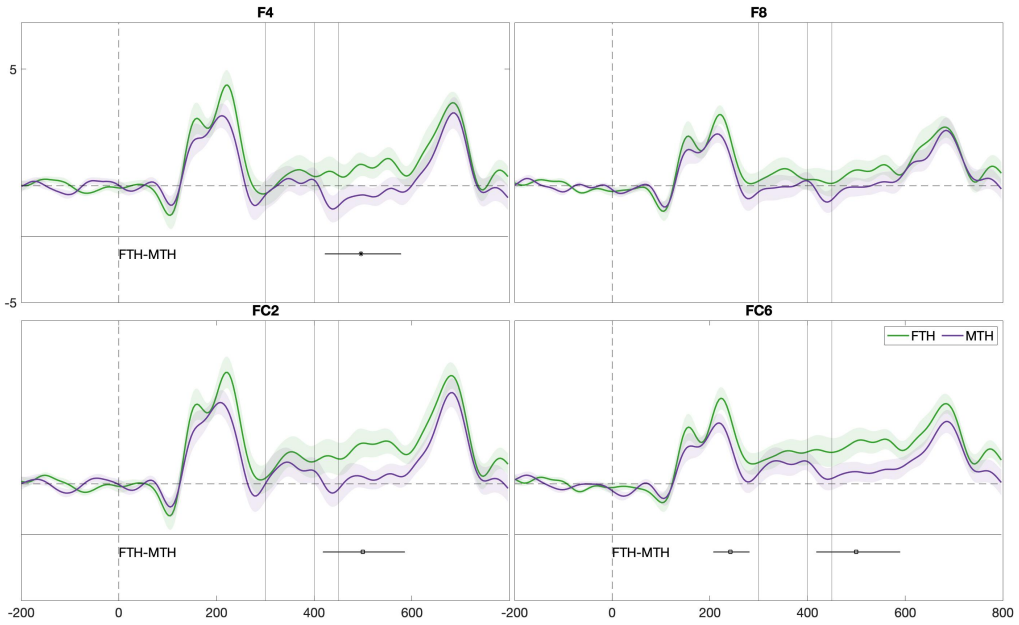
Our ERP analysis replicated the finding of Augurzky et al. (2020). We also found greater late positivity for *fewer than half* than *more than half* between 450 and 800 ms after the stimuli onset. The effect was reflected in a higher EEG amplitude for the negative quantifier (see Figure 5.2). We observed this effect in all regions of interest, however, the difference was more prominent on the centro-parietal

electrodes. No effect of quantifier was found between 300 and 400 ms after the stimuli onset.

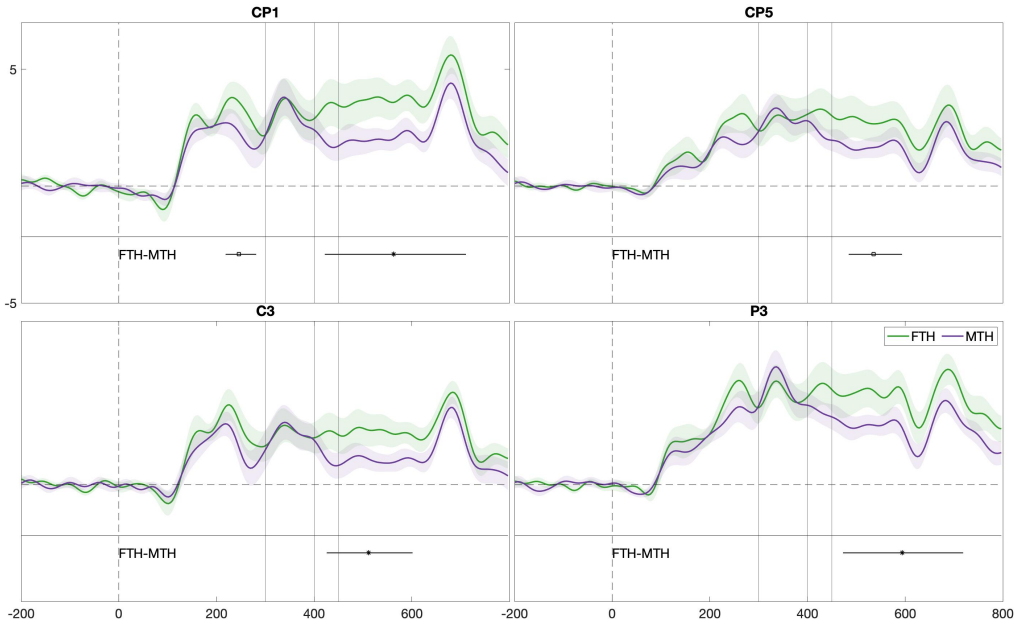
In addition to the replicated late positivity, we also found a difference in EEG amplitude around 200 ms in three ROIs after the stimuli onset. The amplitude was higher for *fewer than half* than for *more than half*, which could reflect the difference in P200 potential between quantifiers (see Figure 5.2).



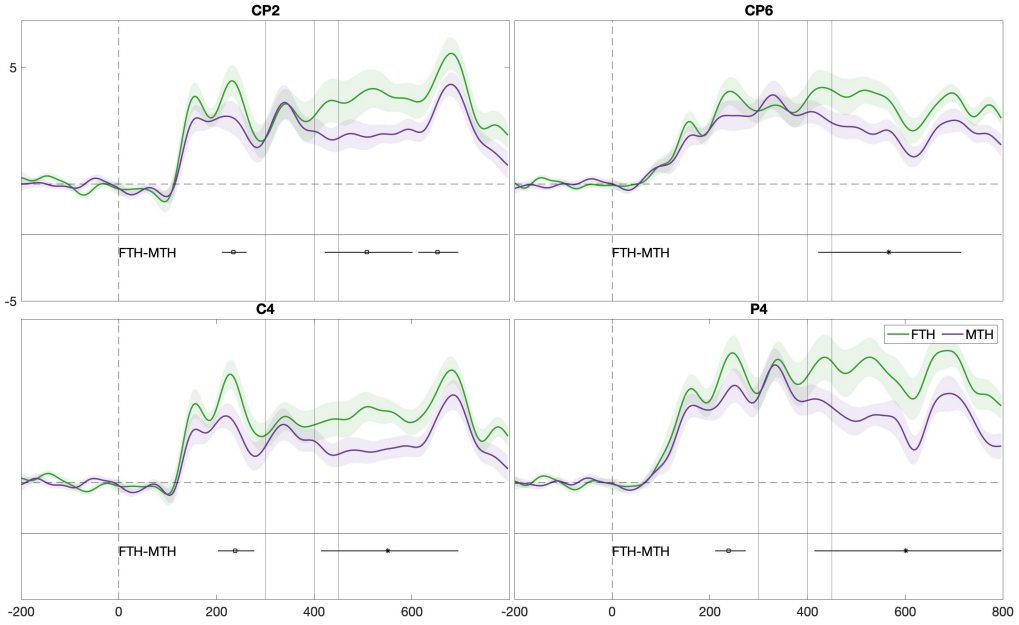
(a)



(b)



(c)

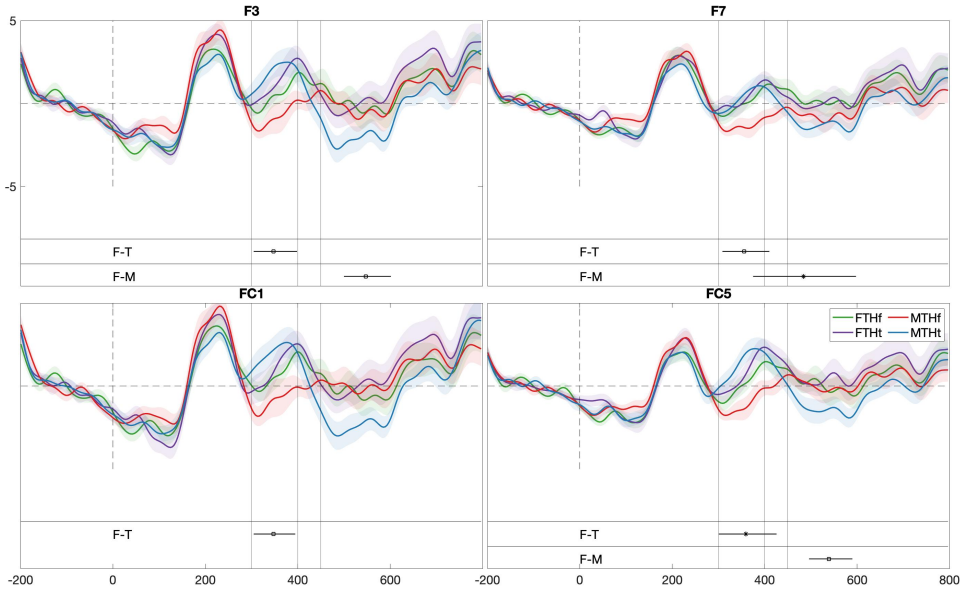


(d)

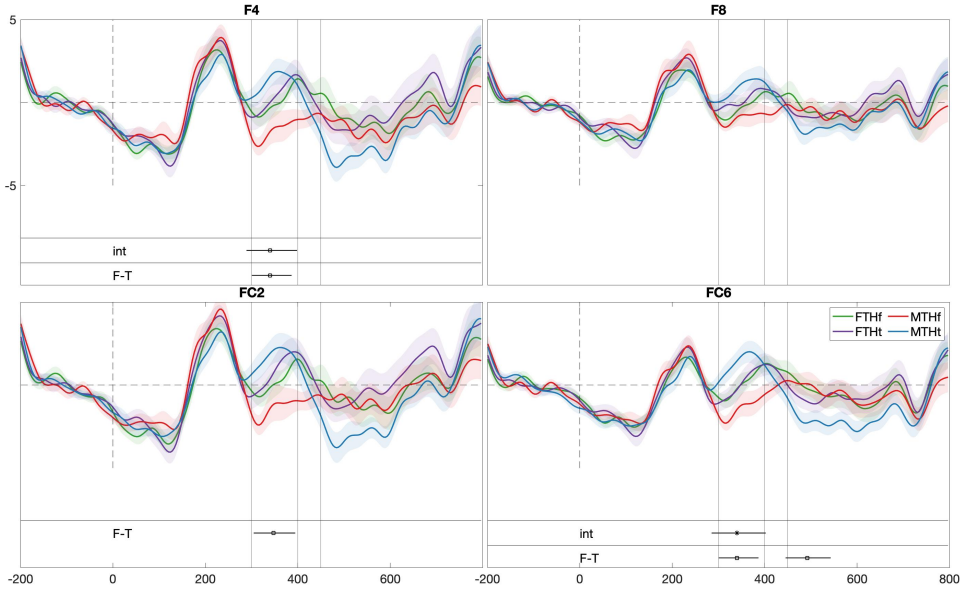
Figure 5.2: ERPs after the quantifier onset in four regions of interest: 5.2a ROI 1, 5.2b ROI 2, 5.2c ROI 3, 5.2d ROI 4. All EEG amplitudes are shown with SE. *Fewer than half* is abbreviated as FTH and *more than half* as MTH. The vertical lines indicate the 300 to 400 ms time window and the 450 to 800 ms time window. The horizontal lines indicate the significant differences, the star indicates the significance level of $p < 0.01$, and the square $p < 0.05$.

After the adjective onset

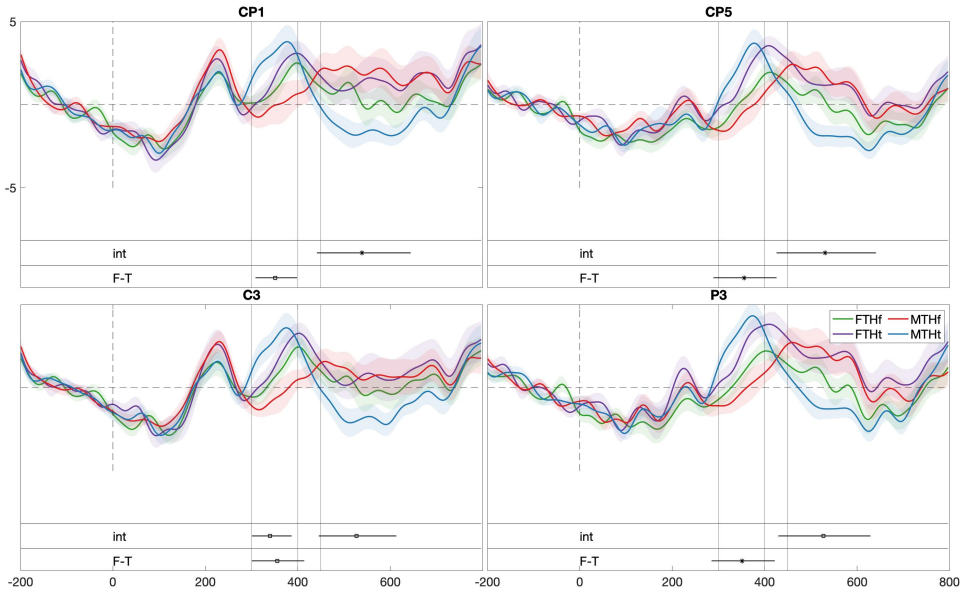
In ERP analysis in the time window after the adjective onset, we found interaction effect between Quantifier and Truth value between 300 to 400 ms in three regions of interest (the interaction was insignificant only in ROI 1, see Figure 5.3). This finding shows the greater negative potential for *more than half* false sentences compared to *more than half* true sentences. We therefore replicated the N400 effect found by Augurzky et al. (2020). In addition, we found a main effect of Truth value in all ROIs between 300 and 400 ms after the adjective onset.



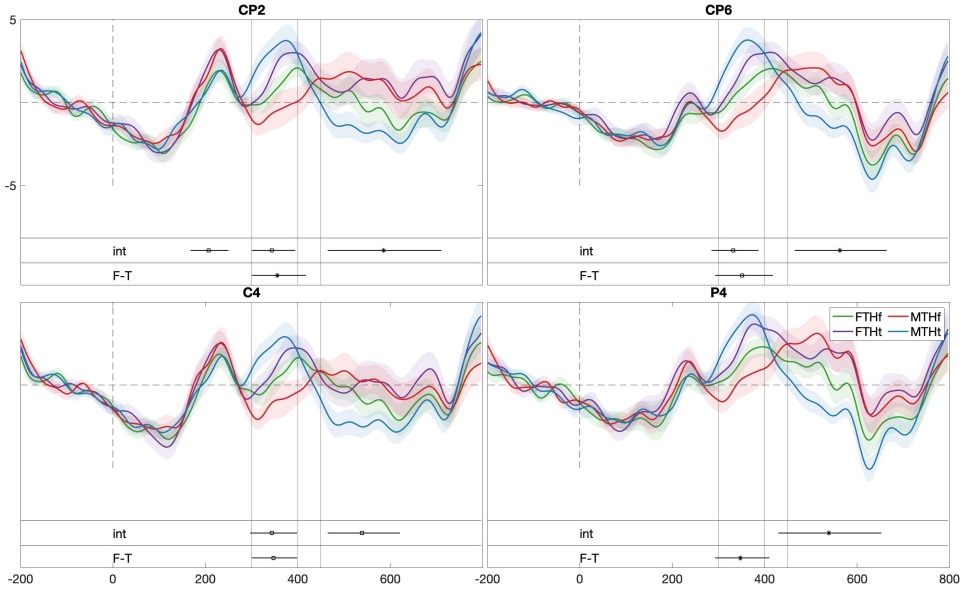
(a)



(b)



(c)



(d)

Figure 5.3: ERPs after the adjective onset in four regions of interest: 5.3a ROI 1, 5.3b ROI 2, 5.3c ROI 3, 5.3d ROI 4. All EEG amplitudes are shown with SE. The vertical lines indicate the 300 to 400 ms time window and the 450 to 800 ms time window. The horizontal lines indicate the significant differences, the star indicates the significance level of $p < 0.01$, and the square $p < 0.05$. The abbreviations indicate: int the effect of interaction, the F-M the effect of Quantifier, the F-T the main effect of Truth value, FTHf *fewer than half* false sentences, FTHt *fewer than half* true sentences, MTHf *more than half* false sentences, and MTHt *more than half* true sentences.

Moreover, we also found an interaction effect in ROIs 3 and 4 in the later time window between 450 and 800 ms after stimuli onset. The EEG amplitude was lower for *more than half* true sentences compared to *more than half* false sentences. In addition, we found a main effect of Quantifier in the same time window in ROI 1 and the effect of the Truth value in ROI 2.

Finally, we also found an interaction effect around 200 ms after the adjective onset in the ROIs 3 and 4. This effect was not reported previously by Augurzky et al. (2020).

ERPs discussion

To summarize, we replicated the Augurzky et al. (2020) findings for short sentences. In the time window after the quantifier onset, we replicated the late negativity effect and also showed the P200 effect. In the time window after the adjective onset, we found an interaction effect between the sentence Truth value and Quantifier between 300 to 400 ms and between 450 and 800 ms. It is worth mentioning that we replicated these effects using a different statistical analysis[11] that controls well for the false alarm rate and limits the changes for the false positive result (Maris & Oostenveld, 2007).

Together with the reaction time results from the previous section, the ERP results encourage the stages of processing analysis. The ERP analyses revealed processing differences between quantifiers that we aim to explain by showing the differences in the processing stages between conditions.

5.3.3 HsMM-MVPA

After the quantifier onset

We fitted the HsMM-MVPA to two conditions (*more than half*, *fewer than half*) separately[12]. The LOOCV analysis revealed that for *more than half* the model with 8 bumps (9 stages) had the highest mean log-likelihood (LL = -213.244). This model had improved fit for a significant number of participants (17 out of 21 participants, sign test $p < 0.05$) compared to the model with 7 bumps (see Appendix C Figure C.4b). For *fewer than half* the results were not unequivocal. The model with 8 bumps had improved fit for only 11 out of 21 participants and did not significantly outperform the model with 7 bumps as indicated by a sign test (sign test $p > 0.05$). However, we are still inclined to select a model with 8 bumps due to mean log-likelihood value (8 bumps LL = -193.672; 7 bumps LL = -193.696; see Appendix C Figure C.4a). Together, this finding does not support the hypothesis that *fewer than half* has more processing stages than *more than half*.

[11] See Appendix C for the ANOVA replication.

[12] See Appendix C for analysis of 500 ms time window.

Because the modeling solution for *fewer than half* was ambiguous between 7 and 8 bumps, we ran an additional analysis in which we fitted one HsMM-MVPA model to the combined data from both quantifiers (combined model). We found that the 8-bumps model was better than the 7-bumps model for 18 out of 21 participants (sign test $p < 0.05$), and it had the highest mean log-likelihood (LL = -394.585) out of all models (see Appendix C Figure C.5). For further evaluation of the differences in the processing stage durations and bumps, we used the 8-bumps model.

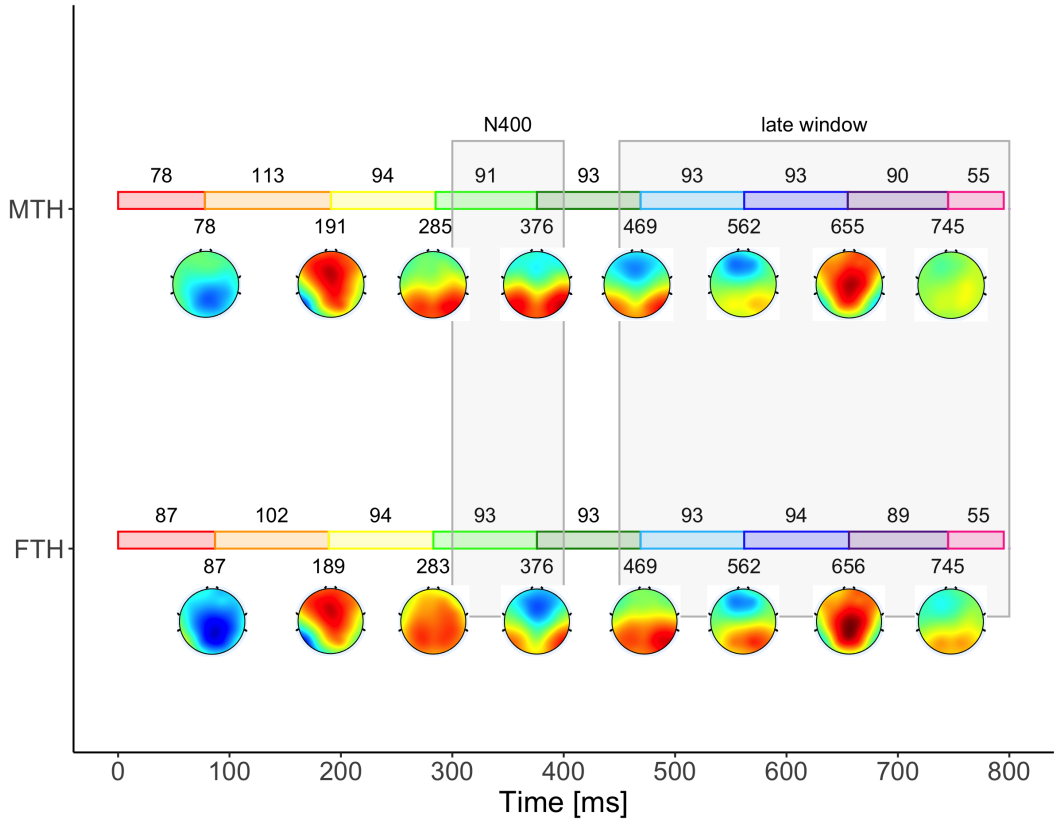
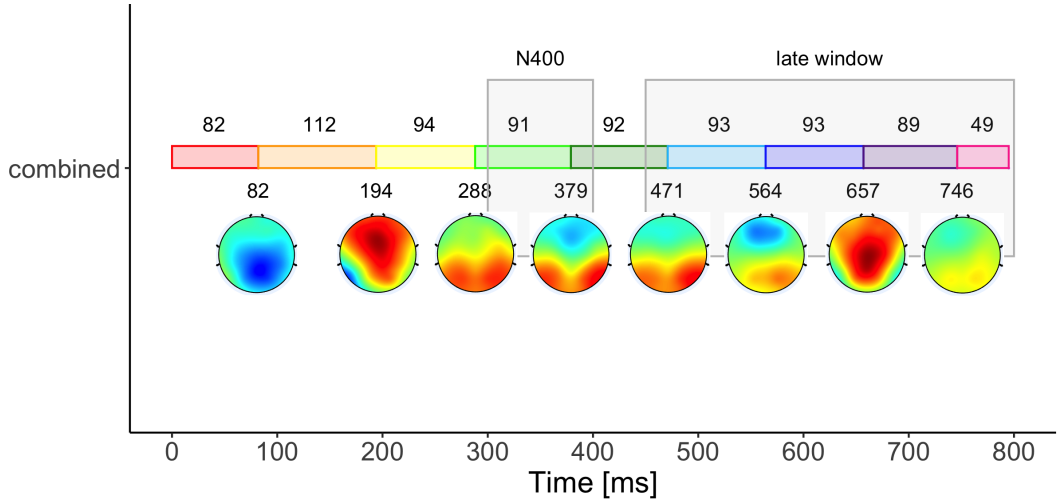
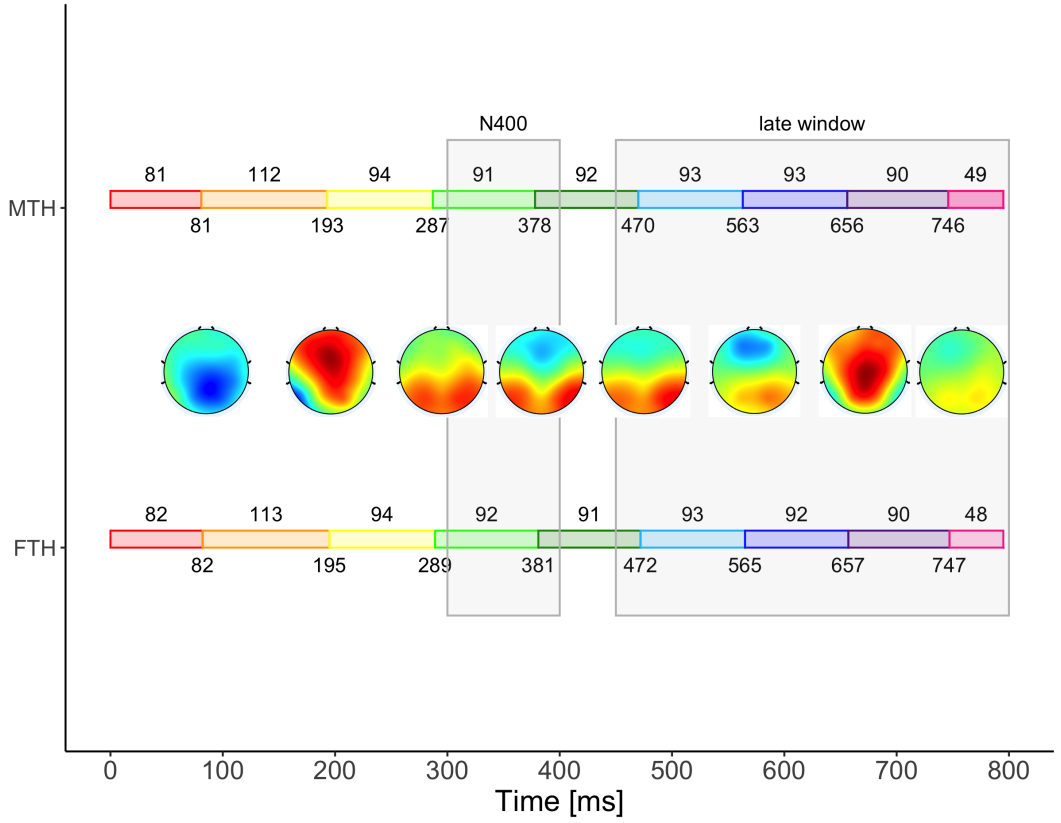


Figure 5.4: Bump topographies and stage durations for separate models (*more than half* (MTH) and *fewer than half* (FTH)) after the quantifier onset. The values above bump topographies correspond to the average onset of the bump. The colored bars indicate the stage durations. The values above the colored bars show the mean stage durations. Additionally, the gray lines indicate the ERP analysis time windows from Augurzky et al. (2020).



(a)



(b)

Figure 5.5: Bump topographies and stage durations for combined model after the quantifier onset. Figure 5.5a shows the stage durations for both conditions together and Figure 5.5b shows the stage durations in each condition separately (*more than half* (MTH) and *fewer than half* (FTH)). The colored bars indicate the stage durations. The values above the colored bars show the mean stage durations and the values below the average onset of the bumps. Additionally, the gray lines indicate the ERP analysis time windows from Augurzky et al. (2020).

Figure 5.4 presents the topographies and average stage duration for 8-bump (9-stage) model fitted to quantifiers separately, and Figure 5.5 shows the topographies and average stage duration for the 8-bump (9-stage) combined model.

After the adjective onset

We followed the same model comparison procedure as in the first time window. Firstly, we fitted the HsMM-MVPA to four conditions *more than half* true sentences, *more than half* false sentences, *fewer than half* true sentences, and *fewer than half* false sentences separately. We found much greater variation in the model fit in all conditions (see Appendix C Figure C.6). For all conditions the model with 10 bumps (11 stages) had the highest log-likelihood: *fewer than half* false LL = -77.0794, *fewer than half* true LL = -72.4346, *more than half* false LL = -75.4606, and *more than half* true LL = -40.9431.

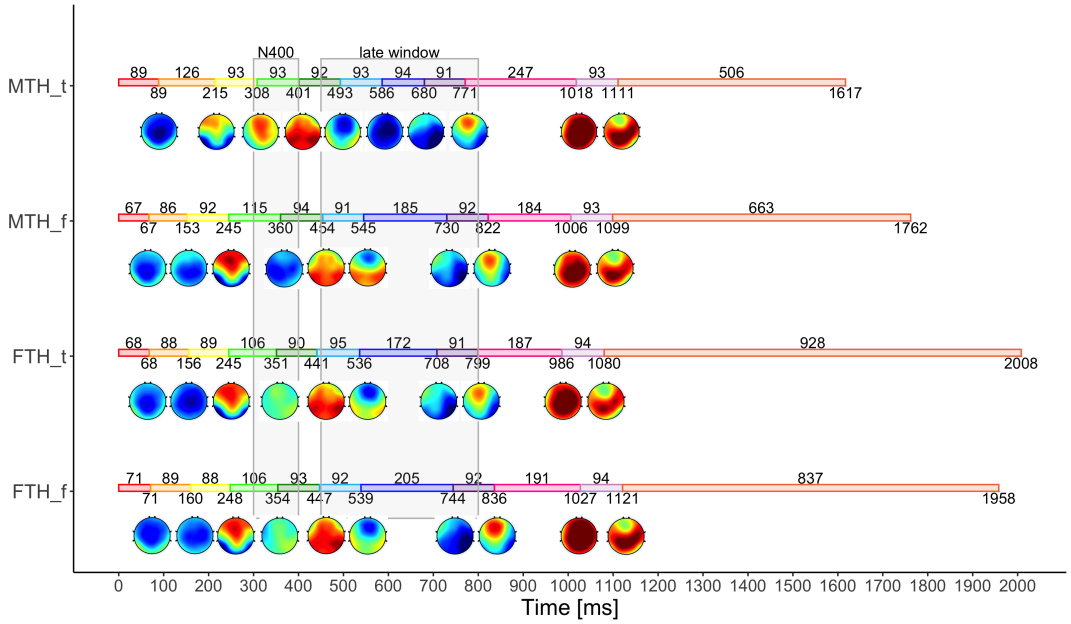
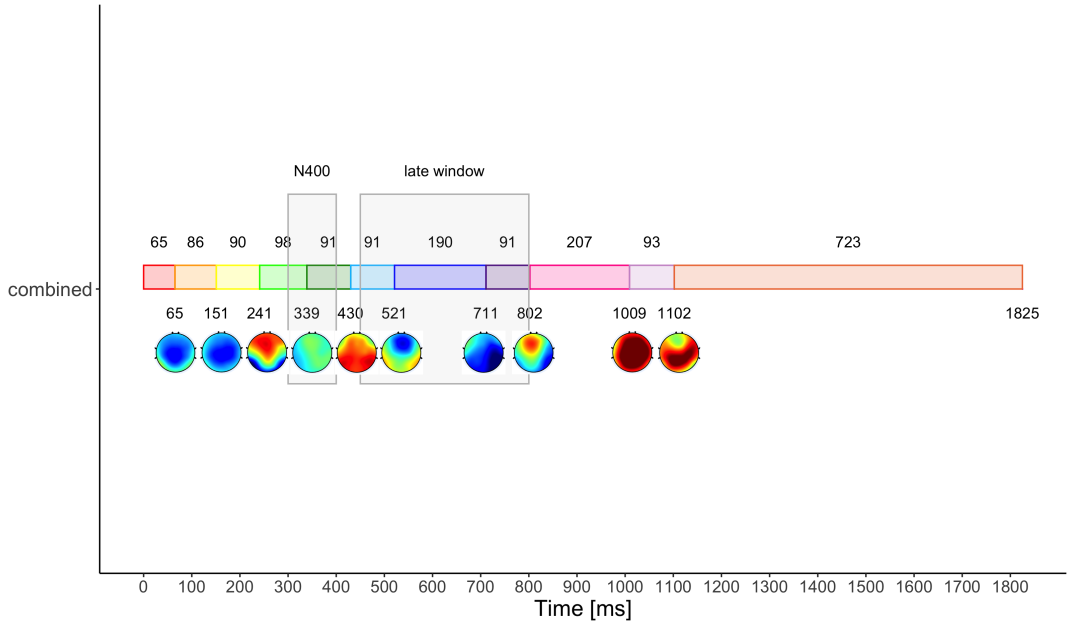
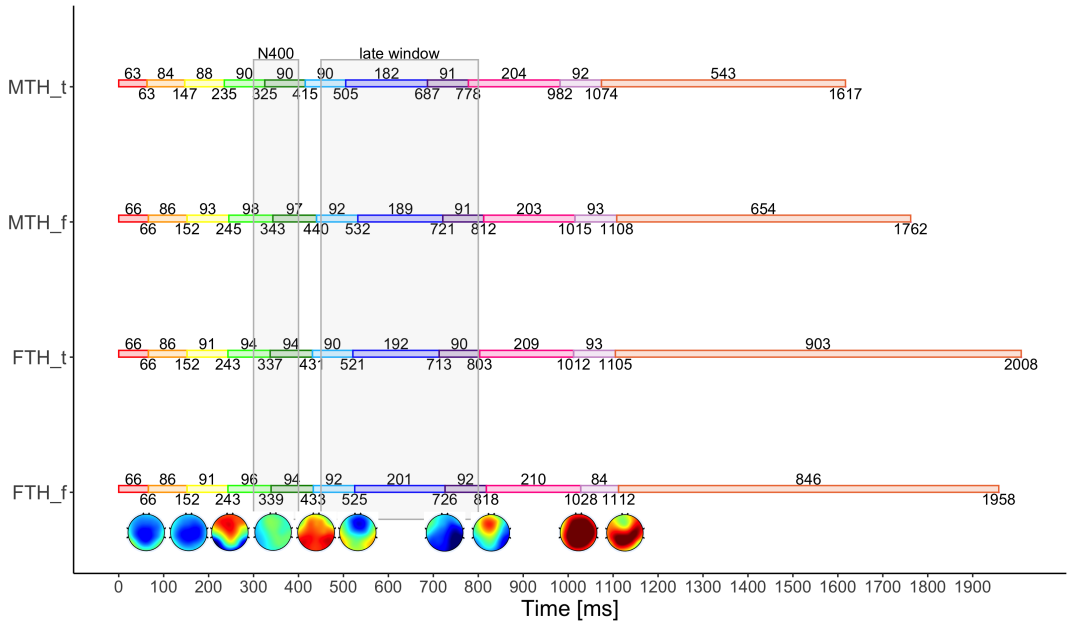


Figure 5.6: Bump topographies and stage durations for separate models (*more than half* true sentence (MTH_t), *more than half* false sentence (MTH_f), *fewer than half* true sentence (FTH_t), and *fewer than half* false sentence (FTH_f)) after the adjective onset until the response. The values above bump topographies correspond to the average onset of the bump. The colored bars indicate the stage durations. The values above the colored bars show the mean stage durations. Additionally, the gray lines indicate the ERP analysis time windows from Augurzky et al. (2020).



(a)



(b)

Figure 5.7: Bump topographies and stage durations for the combined model after the adjective onset until the response. Figure 5.7a shows the durations for all conditions combined and Figure 5.7b separately for each condition (*more than half* true sentence (MTH_t), *more than half* false sentence (MTH_f), *fewer than half* true sentence (FTH_t), and *fewer than half* false sentence (FTH_f)). The colored bars indicate the stage durations. The values above the colored bars show the mean stage durations and the values below the average onset of the bumps. Additionally, the grey lines indicate the ERP analysis time windows from Augurzky et al. (2020).

In the next step, we fitted a combined model to four conditions together. Because the modeling results were not unequivocal, as in the first time window, we wanted to test whether the 10-bump model would fit to all conditions equally well.

The combined model disambiguated the results after the adjective onset. The 10-bump model with mean log-likelihood of LL = -246.924 fitted the data best for a significant number of participants (19 out of 21, sign test $p < 0.05$, see Appendix C Figure C.7). The 10-bump combined model after the adjective onset was better for 15 out of 21 participants than the separate models (sum of mean LL = -265.9177), meaning that the more complex, separate models did not outperform the combined, simpler model. We plotted the bump topographies and stage durations of the separate 10-bump models in Figure 5.6 and of the combined model in Figure 5.7.

HsMM-MVPA discussion

The HsMM-MVPA in both time windows did not support the two-step model prediction. We did not find evidence for the extra processing step for *fewer than half* compared to *more than half*. In both time windows, the more parsimonious combined model outperformed the more complex, separate models. For exploratory purposes, we further investigated the differences in stage durations between *fewer than half* and *more than half*. Our reaction time analysis indicated that *fewer than half* was verified slower than *more than half* and we predict that this difference should be reflected in the processing stages.

5.3.4 HsMM-MVPA mapped models

Our modeling results did not support the hypothesis that *fewer than half* has more processing stages than *more than half* in either of the time windows. The combined models outperformed the separate models in both analyses, see Tables C.1 and C.2.

In the next step, for exploratory purposes, we decided to fit several more complex models with differences between conditions in durations of specific stages and bumps. To test differences in specific stages or bumps, we used so-called mapped models. With mapped models, specific assumptions can be made about stages and bumps. For example, we can construct a mapped model that would assume that the first stage differs between conditions and all the other stages are the same. Similarly, we can define a different mapped model that would assume that the first bump varies between conditions and all the other bumps are shared. In addition, we can also combine these assumptions in the third mapped model where all stages and bumps would be the same across conditions except the first stage and the first bump.

After the quantifier onset

We varied all stages, except Stage 9 (because it was limited by the 800 ms time window) across conditions and bumps. In addition, we tested models in which both stages and corresponding bumps differed between quantifiers. We selected Bumps 3, 5, 6 and 7 and corresponding Stages 4, 6 7 and 8 based on the time windows from Augurzky et al. (2020). None of the more complex models significantly outperformed the simpler combined model. Table C.1 in Appendix C presents the model comparison after the quantifier onset.

After the adjective onset

Following the same procedure used in the analysis after the quantifier onset, we also tested the mapped models with different stage duration and bumps across conditions (see Appendix C, Table C.2). To determine which stages differ between conditions, we ran eleven mapped models, each of which assumed that one stage was different across conditions. None of these models, however, outperformed the combined model. We also tested hypotheses about differences in stage durations and bumps between specific conditions. We derived these hypotheses from Figures 5.6 and 5.7b. We tested models in which: Stages 4 and 5 differed between *more than half* false sentences and were the same for other conditions; Stage 7 differed between *fewer than half* false sentences, *more than half* true sentences and was the same for other conditions; Stage 9 differed for *more than half* true sentences and was the same for other conditions; and Stage 11 differed between *fewer than half* and *more than half*; or *more than half* true sentences, *more than half* false sentences and was the same for *fewer than half* true and false sentences. None of these models outperformed the combined model.

Finally, we also tested the differences in bumps between conditions. We ran mapped models for each bump, assuming that it could have had a different topography across conditions. None of the models, however, outperformed the combined model. In the final model, we tested the hypothesis that the bump was different for *more than half* true sentences and *more than half* false sentences, and the same for *fewer than half* true and false sentences. We tested this bump because its topography seemed to differ the most between conditions (see Figure 5.6), and because it fell in the N400 time window. This model also did not outperform the combined model.

HsMM-MVPA mapped models discussion

The mapped model analyses fail to show differences in bumps or stage durations between conditions. From the analysis after the quantifier onset, we concluded that the onsets of the processing stages were the same across conditions. Moreover, the bumps reflected the same cognitive processes onsets for both quantifiers. The analysis showed that the difference in EEG amplitude between *fewer than*

half and *more than half* does not indicate a delay in the onsets of the consecutive processing stages. We will discuss the methodological implication of this finding in the General Discussion.

The results of the mapped models in the time window after the adjective onset were somewhat surprising. The reaction time analysis indicated that the *fewer than half* was verified slower than *more than half*. This difference was neither explained by the extra processing step nor by longer stages duration in the mapped models. We considered that the mapped models could not have outperformed the combined model due to the small number of trials per condition. Therefore, we further explored the differences in stage duration based on the stages derived from the combined model and tested whether they could predict the reaction times.

5.3.5 Do stages predict the length of reaction times?

In the next step of the analysis, we wanted to test whether the reaction time differences between conditions could be predicted by the stages duration. We used the stage durations estimated from the best fitting model, namely, the combined model (see Figure 5.7).

Fewer than half **false**

For *fewer than half* false sentences, we did not include the by-subject random slope for trial ($\chi^2(1) = 0.73, p = 0.39$). The final model included Stage 4 ($\beta = 0.35, t = 2.33, p = 0.02$), Stage 7 ($\beta = 0.29, t = 3.90, p = 0.0001$), Stage 9 ($\beta = 0.40, t = 5.34, p < 0.0001$), and Stage 11 ($\beta = 0.34, t = 14.64, p < 0.0001$) as significant predictors of reaction times. The intercept of the model was not significant ($\beta = -1.42, t = -1.80, p = 0.07$). The reaction times for *fewer than half* false sentences were thus predicted by Stages 4, 7, 9, and 11.

Fewer than half **true**

For *fewer than half* true sentences, we included the by-subject random slope for trial ($\chi^2(1) = 20.20, p < 0.001$). The final model included Stage 6 ($\beta = 0.89, t = 4.97, p < 0.0001$), Stage 7 ($\beta = 0.15, t = 2.20, p = 0.03$), Stage 9 ($\beta = 0.39, t = 6.94, p < 0.0001$), and Stage 11 ($\beta = 0.35, t = 18.69, p < 0.0001$) as significant predictors of reaction times. The intercept of the model was also significant ($\beta = -3.03, t = -3.91, p = 0.0001$). The reaction times for *fewer than half* true sentences were thus predicted by Stages 6, 7, 9, and 11.

More than half **false**

For *more than half* false sentences, we included the by-subject random slope for trial ($\chi^2(1) = 10.73, p = 0.001$). The final model included Stage 6 ($\beta =$

$0.51, t = 3.20, p = 0.001$), Stage 7 ($\beta = 0.43, t = 6.81, p < 0.0001$), Stage 9 ($\beta = 0.33, t = 5.84, p < 0.0001$), Stage 10 ($\beta = 0.29, t = 1.99, p = 0.047$), and Stage 11 ($\beta = 0.32, t = 17.10, p < 0.0001$) as significant predictors of reaction times. The intercept of the model was also significant ($\beta = -3.58, t = -3.90, p = 0.0001$). The reaction times for *more than half* false sentences were thus predicted by Stages 6, 7, 9, 10, and 11.

More than half true

For *more than half* true sentences, we included the by-subject random slope for trial ($\chi^2(1) = 11.08, p = 0.001$). The final model included Stage 2 ($\beta = 0.49, t = 2.85, p = 0.004$), Stage 4 ($\beta = 0.46, t = 3.09, p = 0.002$), Stage 7 ($\beta = 0.18, t = 2.53, p = 0.01$), Stage 9 ($\beta = 0.56, t = 9.02, p < 0.0001$), and Stage 11 ($\beta = 0.390, t = 16.63, p < 0.0001$) as significant predictors of reaction times. The intercept of the model was also significant ($\beta = -4.60, t = -4.96, p < 0.0001$). The reaction times for *more than half* true sentences were thus predicted by Stages 2, 4, 7, 9, and 11.

5.3.6 Stage durations analysis

In the final step of the analysis, we wanted to test whether the differences in stage duration between quantifier and truth value conditions. We found the significant main effect of Quantifier and interaction in the reaction times data. We expected that this effect should be reflected in the duration of stages. Because the combined model turned out to be the best fitting model, we decided to run an additional analysis on the stage durations extracted from this model. We tested stages that were significant predictors of reaction times for all experimental conditions, namely Stages 7, 9, and 11. We plotted the stage durations from the combined model separately for each condition in Figure 5.7b.

Stage 7

Firstly, we tested the differences in Stage 7. We included the by-subject random slope for trial ($\chi^2(1) = 22.49, p < 0.001$). The interaction between Quantifier and Truth value was not significant ($\beta = 0.003, t = 0.11, p = 0.91, \chi^2(1) = 0.01, p = 0.91$). The best model had significant two main effects: of Quantifier ($\beta = -0.03, t = -2.14, p = 0.03$) and Truth value ($\beta = -0.04, t = -2.71, p = 0.007$), and intercept ($\beta = 5.18, t = 191.86, p < 0.001$). We found that *fewer than half* had longer Stage 7 than *more than half*, and that this stage was longer for false sentences compared to true sentences (see Figure 5.8).

Stage 9

Secondly, we tested Stage 9. We included the by-subject random slope for trial ($\chi^2(1) = 25.92, p < 0.001$). We found that the Quantifier x Truth value interaction was not significant ($\beta = 0.006, t = 0.21, p = 0.84, \chi^2(1) = 0.04, p = 0.84$). The model without interaction had a significant intercept ($\beta = 5.25, t = 176.55, p < 0.001$), but neither a main effect of Quantifier ($\beta = -0.02, t = -1.51, p = 0.13$), nor a main effect of Truth value ($\beta = -0.01, t = -0.77, p = 0.44$). We conclude that this stage did not differ between experimental conditions (see Figure 5.8).

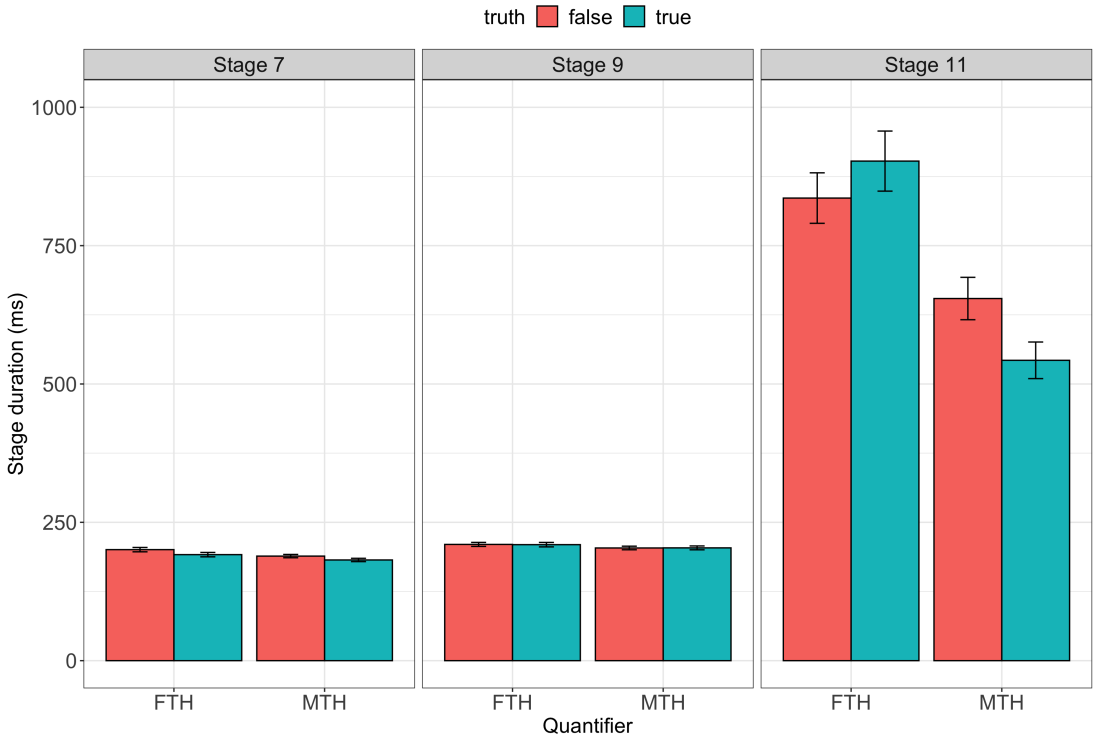


Figure 5.8: Mean durations of Stages 7, 9, and 11. *Fewer than half* is abbreviated as FTH and *more than half* as MTH. The error bars represent within-participant SE.

Stage 11

Next, we tested the last stage, Stage 11[13]. This stage ended when participant provided a response. Therefore, we expected this stage to differ in a similar way

[13]We observed the same problem with Stage 11 distribution as with reaction time distribution. The stage of the distribution was binomial even after log-transformation and, therefore, the model did not fit the assumption of normal distribution of residuals. We ran additional analysis on the data without timeout reaction times in Appendix C

as reaction times.

We included the by-subject random slope for trial ($\chi^2(1) = 92.14, p < 0.001$). We found that the Quantifier x Truth value interaction was not significant ($\beta = -0.10, t = -1.24, p = 0.22, \chi^2(1) = 1.52, p = 0.22$). The model without interaction had a significant intercept ($\beta = 5.80, t = 42.56, p < 0.001$), and a main effect of Quantifier ($\beta = -0.23, t = -5.47, p < 0.001$) and insignificant main effect of Truth value ($\beta = -0.04, t = -1.05, p = 0.30$). We found that *fewer than half* had a longer Stage 11 than *more than half* (see Figure 5.8).

Combined effect of Stages 7, 9, and 11

Finally, we summed the durations of Stages 7, 9, and 11, to test whether their combined duration could be predicted by the effect of Quantifier and Truth value. We found that the random slope for trial was significant ($\chi^2(1) = 126.08, p < 0.001$). The interaction effect was not significant ($\chi^2(1) = 1.36, p = 0.24$). After excluding the interaction from the model, we found that the main effect of Quantifier ($\beta = -0.15, p < 0.001$) and intercept ($\beta = 6.70, p < 0.001$) were significant, but the main effect of Truth value was not ($\beta = -0.04, p = 0.10$).

5.4 General Discussion

The main goal of the current study was to test whether the two-step model prediction that negative quantifiers require an extra processing step compared to positive ones. The extra step could be related to the higher complexity of the representation of negative quantifiers or to a longer verification procedure (e.g., Clark, 1976; Grodzinsky et al., 2018). To test this hypothesis, we analyzed data from the quantifier picture-sentence verification task collected by Augurzky et al. (2020). We used a novel HsMM-MVPA method (Anderson et al., 2016) to detect the stages of processing in EEG signal and directly compared them between quantifiers. We analyzed two time windows in which Augurzky et al. (2020) previously found a difference in EEG amplitude between quantifiers. Our analysis did not support the two-step model. We found the same number of stages in all experimental conditions in both time windows. In the next sections of the discussion, we summarize our main findings and propose an interpretation of our results in light of an alternative account. Moreover, we elaborate on the methodological implications of our study.

5.4.1 After the quantifier onset

Our finding in the time window up to 800 ms after the quantifier onset did not support the extra step of the processing hypothesis. Firstly, both separate models favored the 8 bumps solution, although the result was weaker for *fewer than half*.

Secondly, we found that the 8 bumps model with combined conditions of both quantifiers fitted the data better than the separate models.

Moreover, we also found that the onsets of bumps were very similar across conditions and that there was little variation in stage durations between quantifiers. We tested various models assuming differences in each stage duration. However, none of these models outperformed the combined model. This finding led us to the conclusion that the time course of processing the quantifier at the beginning of the sentence is fixed and the onsets of cognitive processes for *fewer than half* were not delayed.

Nonetheless, in the ERP analysis[14], we found a greater late positivity for *fewer than half* than *more than half*. This finding replicated the previous result of Augurzky et al. (2020). Together, our findings showed that the onsets of the upcoming cognitive processes were not delayed in *fewer than half*, but they could have been more cognitively costly. Two factors could explain the difference in EEG amplitude between quantifiers. Firstly, the HsMM-MVPA detects the onsets (Anderson et al., 2016) of cognitive stages, but not the offsets. In principle, the next cognitive process could have started before the previous one was finished. If two or more cognitive processes were ongoing in parallel, then their combined effect could have resulted in a larger EEG amplitude. According to this explanation, the stages of processing for *fewer than half* overlapped, which led to a higher EEG amplitude compared to *more than half*. In a line with the alternative explanation, the cognitive processes could have finished at the same time for both quantifiers, but for *fewer than half* they were more costly and required engagement of more neural resources. In the next section of the discussion, we propose an interpretation of the cognitive stages of the combined model.

Functional interpretation

Following previous HsMM-MVPA studies (e.g., Anderson et al., 2016; Berberyan et al., 2021), we identify Stage 1 as a pre-attention stage. In our study, this stage had an average duration of 8 2ms, which is consistent with the analysis by Anderson et al. (2016) and the prediction from a theoretical model (ACT-R, 85 ms, Anderson et al., 2016). Moreover, we found a similar brain activation in the pre-attention stage to that found in the Anderson et al. (2016) study.

The second stage started with the onset of the negative bump. We identify this stage as the visual encoding stage (Anderson et al., 2016; Berberyan et al., 2021). The negative bump is likely to correspond to the N100 visual potential (Luck, Woodman, & Vogel, 2000), with negativity distributed mostly on parietal regions. The second stage in our study took 112 ms on average, which is comparable to the Anderson et al. (2016) study (95 ms).

Stage 3 started with the fronto-centroparietal positive bump, which could have

[14]See Appendix C for replication of ANOVA analysis.

reflected the P200 potential. We found that the amplitude of P200 was higher for *fewer than half* than *more than half* on some electrodes. A similar stage reflecting P200 potential was found in another linguistic task with HsMM-MVPA (Berberyan et al., 2021). The second bump also has similar onset to the typical reading time of the word, namely around 200 ms (Swaab, Ledoux, Camblin, & Boudewyn, 2011). The P200 effects were observed in the word-by-word sentence reading (Dambacher, Kliegl, Hofmann, & Jacobs, 2006). Moreover, the timing of the second bump and Stage 3 are similar to lexical retrieval predicted by the memory retrieval model of Anderson et al. (2016). Our HsMM-MVPA analysis showed that this stage had an average duration of 94 ms, which is almost identical to the third-stage duration in the Anderson et al. (2016) study (95 ms). We thus assumed that, in the third stage, participants have already read the displayed word and we identified Stage 3 as the lexical access stage.

Stage 3 was followed by Bump 3 with a positivity distributed on parietal regions. Given the timing of Stage 4 and the topography of the Bump 3, we hypothesized that this stage reflects the P300 component. P3a and P3b constitute a broader class of P300 components (Polich, 2007), and both are involved in the evaluation of a new stimulus. P3a is typically related to the monitoring of attention, especially in the presence of a distractor. It has a more frontal distribution. P3b, in contrast, has a parietal distribution and reflects context updating and maintenance. In the HsMM-MVPA study, Berberyan et al. (2021) found the attention stage accompanied by P300 in the lexical decision task. Moreover, van Maanen et al. (2021) found a positive bump with a posterior distribution in a speed-accuracy trade-off task in a focused condition. We also linked Stage 4 with the engagement of attention resources in the processing of the first word of the sentence.

An alternative interpretation of Stage 4 would link it with a decisional account of P300. The P3b component is a signature of the evidence accumulation process (Twomey, Murphy, Kelly, & O'Connell, 2015). It reflects the process of reaching the decision boundaries. In our experiment, the evidence accumulation was gradual and progressed with each new word revealed. It was stretched through the whole process of sentence comprehension. Its neural signatures could already have occurred at the beginning of the sentence processing.

Stage 5 has a somewhat similar bump topography to Stage 4, with a difference in frontal negativity. Based on the previous HsMM-MVPA studies (Anderson et al., 2016; Berberyan et al., 2021), we linked Stage 5 with the familiarity-driven recognition. The bump signaling the onset of the familiarity-driven recognition stage also has a frontal negativity characteristics.

Alternatively, Stage 5 can also reflect the early N400 potential (Kutas & Hillyard, 1980; Kutas & Federmeier, 2011). The N400 interpretation, however, does not fit the topography of Bump 4. The typical N400 component has centroparietal distribution, while Bump 4 has a rather frontal distribution. Moreover, in Stage 5, we can observe more positive activity in parietal regions.

Stage 6 begins with a parietal distributed positive bump that may reflect the late positivity found by Augurzky et al. (2020) and replicated in the current study. The increasing positive activity can reflect the P600 potential often found in the linguistic tasks (see Brouwer et al., 2012, 2017, for review). In our study, the positive parietal activity is characteristic for four consecutive stages (Stages 4 to 7). This observation is consistent with literature showing multiple functional interpretations of P600, probably linked to different underlying components (Regel, Meyer, & Gunter, 2014). For example, P600 is sensitive to ungrammatical structures as well as pragmatic manipulation (Regel et al., 2014), e.g., the irony. The so-called pragmatic P600 is preceded by the P200 component. Moreover, while the syntactic P600 is widespread over the scalp, the pragmatic P600 is mostly visible on central and parietal electrodes. However, both components have similar latency around 50 0ms after the stimuli onset. We found a pattern of results characteristic to the pragmatic P600: a large peak of P200 component with a significant difference between quantifiers in Stage 3 (see the ERP analysis in Figure 5.2 and Figure 5.5 for stages onsets), and a P600 difference in the time window of Stage 6. We therefore suggest that Stage 6 can be interpreted as a stage of processing pragmatic properties of quantifiers. Moreover, we suspect that Stage 6 may have continued after the onset of the next stage, because we can still observe the difference in amplitude between quantifiers on parietal electrodes (see the ERP analysis in Figure 5.2 and Figure 5.5 for stages onsets).

Stage 7 starts with a characteristic bump with frontal negativity and posterior positivity distribution. The frontal negativity could reflect the engagement of working memory. The negative slow wave (NSW) has a frontal distribution in verbal tasks (Ruchkin, Johnson, Grafman, Canoune, & Ritter, 1992). For example, Bailey, Mlynarczyk, and West (2016) found in typical working memory (N-back task) that the slow negative wave was associated with maintaining information in the working memory. At the onset of Stage 7, the new word was already present on the screen and participants started building the sentence structure. The new word might have increased the load in their working memory.

Alternatively, the negative frontal bump could reflect the late N400 effects. The typical N400 potential has an onset around 300 to 400 ms after the stimuli and reflects the semantic processing (Kutas & Hillyard, 1980; Kutas & Federmeier, 2011). However, it can also have later timing, around 500 ms, and frontal distribution. For example, the frontally distributed N400 was linked to conceptual expectations (Thornhill & van Petten, 2012), and the late frontal negativity to hierarchical relationships (Chen et al., 2014).

The widespread positive activation in Stage 8 can reflect the syntactic processing of the sentence. The topography of the seventh bump matches the syntactic P600 description (Regel et al., 2014). Moreover, the bump occurs after the onset of the next word on a screen, which could reflect syntactic integration of the upcoming word. In addition to syntactic integration, it could also reflect the integration of contextual information (Brouwer et al., 2017).

We do not give any interpretation of Stage 9 because this stage was not finished due to the time window limit. Finally, we noticed that the duration of the consecutive cognitive stages (except Stages 1 and 2 and the last stage) was very similar, between 90 and 94 ms.

5.4.2 After the adjective onset

In the time window after the adjective, four separate models for each quantifier and truth value combination showed a large variation in model fit across participants. However, the 10-bump model had the highest mean log-likelihood for all conditions. We fitted the combined model to all conditions jointly to establish the most likely number of bumps. The model with 10 bumps fits the data most accurately. The combined model also outperformed the four separate models, meaning that, in the time window after the adjective onset, we did not find support for an extra processing step for negative quantifier. We further analyzed the processing stage of the 10-bump combined model.

The variability in a model fit of separate models could be explained from methodological and theoretical perspectives. Methodologically, the problem with model fit could be a result of an insufficient number of trials per condition. The previous HsMM-MVPA studies analyzed data from experiments in which participants completed at least 100 trials per condition (Anderson et al., 2018)[15]. This explanation seems plausible, especially because we obtained a more consistent model fit across participants for the combined model. From a theoretical perspective, the variation in model fit across participants could reflect individual differences in the processing of quantified sentences. Individual differences were found previously in semantic representations of vague quantifiers (Ramotowska et al., 2020b) and in the different pragmatic interpretations of ambiguous sentences (Spychalska et al., 2016). Given the small number of trials per condition in the current experiment, we can not exclude any of discussed interpretations. Nonetheless, this result calls for further exploration with a larger sample size.

Since we did not find support for the two-step model hypothesis, we further explored the differences in the stage durations between conditions. We found the effect of Quantifier and significant interaction between Quantifier and Truth value in the reaction time data and we predicted that this effect should be reflected in stage durations.

Firstly, we explored the difference in stage durations using the mapped model. However, none of the mapped models however outperformed the combined model. We think that the absence of a significant effect could be due to the small number of trials and the problem with model fit rather than the absence of a true difference in the population. Therefore, we derived the stage durations from the combined

[15]See also the discussion about the sufficient number of trials to estimate parameters of other cognitive models, e.g., Wagenmakers (2009); Lerche et al. (2017); Osth, Dennis, and Heathcote (2017); Boehm et al. (2018).

model for each quantifier and truth value combination and tested how well they could predict the reaction times. We found that three stages (Stage 7, 9, and 11) predicted reaction times across conditions. We tested the differences between quantifiers in these stages. Stages 7 and 11 showed differences between quantifiers and, in addition, Stage 7 showed differences between truth values. However, none of the stages alone nor their combined duration reflected the interaction effect found in reaction time data. This finding was unexpected, but we assume that the interaction effect could also depend on stages that predicted reaction times only for some experimental conditions and which were not included in the regression model. Moreover, we note that the effect in Stage 11 went in the direction of the interaction.

As in the earlier time window, also after the adjective onset, we replicated the ERP effects found by Augurzky et al. (2020)[16]. We found the N400 effect only for *more than half* false sentences as well as later amplitude deflection for *more than half* true sentences.

In the next section of the discussion, we propose the functional interpretation of 11 stages after the adjective onset.

Functional interpretation

The first stage of processing after the adjective onset could be related to the processing of the previous word. Therefore, we do not give any specific interpretation of this stage because we do not know when was its onset occurred, and what its bump topography was.

The next stage started with the negative bump distributed over the whole scalp. This bump possibly reflects the visual encoding stage, similar to the first bump after the quantifier onset. We noticed, however, that the first bump onset was earlier (about 65 ms) in all conditions after the adjective onset than in the time window after the quantifier onset (about 80 ms). Moreover, in the time window after the adjective onset, the first bump was followed by the second, which was also negative bump. Thus, we hypothesized that, in the time window after the adjective onset, the second bump reflects the visual encoding of the adjective (N100), while the first bump could reflect either the processes related to the earlier part of the sentences or the so-called semantic predictive potential (SPP), a slow negative wave before the stimuli onset that indicates the predictions about the upcoming word (Grisoni, Tomasello, & Pulvermüller, 2021; Grisoni, Mccormick Miller, & Pulvermüller, 2017). We noticed that the SPP should appear before the adjective onset. Nonetheless, we consider that the 65 ms was not enough time for participants to process the adjective. Therefore, the first bump could be still related to processes before the final word of the sentence appeared on the screen.

[16]See Appendix C for ANOVA analysis replication.

In line with this interpretation, Stage 2 after the adjective onset had an average duration of 86 ms, which is almost identical to the duration predicted by the theoretical model and HsMM-MVPA in the Anderson et al. (2016) study (85 ms) and very similar to the duration of the first stage after the quantifier onset in the current study (82 ms).

Following this interpretation further, we identified Stage 3 as the visual encoding stage. This stage was shorter during encoding of the adjective (around 90 ms) than the quantifier (110 ms). However, its length may depend on the length of the to be encoded word. In the Anderson et al. (2016) study, the corresponding Stage 2 had a duration of 95 ms. It is worth mentioning that the duration of Stage 3 and the onset of Bump 2 were very similar across conditions, indicating no apparent delay in processing for *fewer than half*.

The fourth stage took around 98 ms and started with a positive bump. The topography of Bump 3 resembles Bump 2 after the quantifier onset. The duration of this stage was also very similar to the duration of the respective stage after the quantifier onset. We identify Stage 4 after the adjective onset as the lexical information retrieval with a prominent P200 component.

The fifth stage after the adjective onset is likely the semantic encoding stage with the N400 component. Our model comparison analysis did not reveal a difference in Bump 4, which most likely indicates N400. The lack of a positive result was somewhat surprising, but may not have been detected due to the methodological reasons mentioned above. Nonetheless, we found the N400 effect in the ERP analysis (see Figure 5.3).

We proposed that the sixth stage is a context update stage with a P600 potential. This stage could reflect the integration process. This finding is consistent with the Retrieval-Integration account of N400 and P600 (Brouwer et al., 2012, 2017). The integration processes typically follow the N400. Interestingly, Stage 6 did not seem to differ in duration between conditions, although its onset was earlier for *more than half* true sentences for about 15 to 20 ms.

The seventh stage started with Bump 6 with a negative frontal distribution. This stage was longer for the negative quantifier than for positive one and for false sentences than true. Therefore, we suggest that this stage is related to a cognitive process that is crucial to the truth value evaluation. The negative frontal bump could reflect the working memory processes (Ruchkin et al., 1992). Participants involved their working memory in the comparison between picture and sentence representations.

The stage of comparison between sentence and picture was described in detail by Clark (1976). The 'true' model of negation predicts the interaction between sentence truth value and polarity. However, the pattern of our result in Stage 7 seems to be more compatible with the 'conversion' model. This model does not predict the interaction in the comparison stage, but rather the main effects of truth value and polarity. Our findings in Stage 7 are in line with this prediction. Although we found an interaction effect in reaction times, this effect

was not present in Stage 7. This finding indicates that the comparison between mean reaction times can be misleading because reaction times are contaminated by effects from different processing stages. The analysis of reaction times and stage duration can lead to different conclusions about the validity of the negation models.

We identified the eighth stage as the truth value evaluation. The truth value evaluation was linked to late negativity between 500 to 1000 ms after the critical word onset (Wiswede et al., 2013). We based this interpretation on the bump topography, which resembles sustained negativity. Sustained negativity was previously found to be related to the pragmatic reinterpretation of scale implicatures (Zhao, Liu, Chen, & Chen, 2015). It reflects the pragmatic effort in reanalyzing the sentence.

The second possible interpretation of Stage 8 is that it reflects the decision process. Previous HsMM-MVPA studies linked the decision stage to a negative bump (Anderson et al., 2016; van Maanen et al., 2021) in the associative recognition task.

The previous HsMM-MVPA studies (Anderson et al., 2016; van Maanen et al., 2021; Berberyan et al., 2020; Zhang, Walsh, & Anderson, 2017) showed that the positive bump is typically related to the response stage. Our analysis revealed three different consecutive stages starting with a positive bump. This split into three separate stages could be due to the experimental procedure, which required participants to hold the response until they saw the signal.

The first stage starting with a positive bump is Stage 9. It starts with a characteristic left frontal positive bump, similar to the late positive complex (LPC). The LPC is related to the inhibition of the response (De Jong, Coles, Logan, & Gratton, 1990). Specifically, frontocentral left positivity can be observed in the go-no-go task in the no-go trials. It indicates the successful inhibition of response (Kiefer, Marzinzik, Weisbrod, Scherg, & Spitzer, 1998). The left distribution of the potential is independent of the response hand (Kiefer et al., 1998), and it is generated by the left premotor cortex. A similar LPC effect was found in the working memory task (Kusak, Grune, Hagendorf, & Metz, 2000), where the positive activity was linked to the central executive in working memory. The frontal LPC was also associated with the general task performance and adaptation to the experimental setup (Pauli, Lutzenberger, Birbaumer, Rickard, & Bourne, 1996). The LPC reaches the maximum on the centroparietal electrodes (Kusak et al., 2000), which resembles the ninth bump in our analysis.

Stage 10 started with a large positive bump spilled over the whole scalp. Its distribution resembles bumps from the previous HsMM-MVPA studies (Anderson et al., 2016; Berberyan et al., 2021). This bump can reflect the preparation for the response. It can also indicate the centroparietal positivity (CPP), which was linked to reaching the decision boundaries in evidence accumulation models (Twomey et al., 2015).

The third positive bump related to the decision processes occurs at the start

of the last stage, Stage 11. This stage ends with participant response. Stage 11 was longer for the positive quantifier than the negative quantifier. However, the Truth value and interaction effects in this stage were not significant. Contrary to previous HsMM-MVPA studies (Anderson et al., 2016; Berberyan et al., 2021), the last stage in our model was the longest. Typically, the decision stage is followed by a short response stage, which includes only the execution of response. For example, according to theoretical model (ACT-R model) predictions by Anderson et al. (2016), the last stage should take around 50 ms to complete motor preparation and 60 ms to the execute the response. The response stage in our experiment could be longer than in previous studies due to the higher complexity of the task. This stage could also be longer because participants made a decision before the onset of the signal to respond and they their attention drifted elsewhere (similar results were obtained by e.g., Kaup et al., 2006).

We found a difference between quantifiers in the two processing stages. This finding is consistent with previous Diffusion Decision Model (DDM, Ratcliff, 1978) findings on quantifiers (Schlotterbeck et al., 2020). In this experiment, Schlotterbeck et al. (2020) found a difference in two DDM parameters, namely the drift rate and non-decision time. Although we can not directly translate our finding to the DDM model, we also found a difference in the stage related to the comparison between representations, presumably part of the decision process, and in the final response stage, which could reflect the difference in non-decision time.

5.4.3 Alternative explanation of the polarity effect

The results of the HsMM-MVPA analysis do not support the two-step model. However, they could be interpreted in light of the competing approach, namely the pragmatic account. As in the case of the two-step model, we use the label 'pragmatic account' to indicate a broader class of proposals that explain the differences between negatives and affirmatives in terms of the speaker's pragmatic preferences and interaction between processed sentence and discourse context.

The pragmatic account found support in EEG studies on negation and quantifiers. Pragmatic information influences the processing of negation more than the processing of affirmative sentences (Orenes et al., 2016). For example, Nieuwland and Kuperberg (2008) showed in the EEG experiment that difficulties in processing negation disappear in the pragmatically licensed context. Moreover, negative quantifiers (such as *few*) can also be processed fully incrementally in an appropriate discourse context (Urbach et al., 2015). Nieuwland (2016) demonstrated that the difficulties of processing quantifiers are dependent on the predictability of the sentence continuation.

Further support for the pragmatic account comes from the computational modeling results. Schlotterbeck et al. (2020) fitted the Diffusion Decision Model (Ratcliff, 1978) to the responses and reaction times data from two verification tasks. They argued that the pragmatic account predicts the difference in drift

rates between positive quantifier (*more than half*) and negative quantifier (*fewer than half*). They found that the evidence accumulation was faster for positive expression. Ramotowska et al. (2020a) replicated this finding.

Two questions remain, however: (1) what types of context do facilitate the processing of negative expressions, (2) and what is the mechanism of context effect? Xiang, Kramer, and Nordmeyer (2020) suggested two possible mechanisms. The informativity-based account (Xiang et al., 2020) claims that the negative sentences are usually underinformative. For example, imagine that you aske your friend: "What did you eat for dinner yesterday?" and the friend answeres: "I did not eat soup." This answer gives you very limited information about your friend's dinner. You would probably be more satisfied with a positive answer. The expectation-based account (Xiang et al., 2020), in turn, proposes that negation is usually unexpected and thus generates the processing cost. The underinformative answer from the previous example seems inappropriate and surprising without any additional contextual information about your friend's eating habits. In the next section, we will evaluate whether these two proposals can account for our results.

Informativity-based account

According to the informativity-based account, negative expressions are dispreferred on the pragmatic grounds because they are less informative (Horn, 2001). Xiang et al. (2020) found support for this account in a behavioral task. In the EEG experiment, Nieuwland, Ditman, and Kuperberg (2010) showed that the participants with higher pragmatic skills are more sensitive to a sentence's underinformativeness.

The informativity-based approach cannot, however, account for our findings. In the Augurzky et al. (2020) experiment, the pictures were constructed in such a way that both quantifiers were equally informative. The sentence that was false for *more than half* was true for *fewer than half* and the other way around. Moreover, the pictures contained shapes in only two colors. Therefore, the information about the color of the greater set was equally informative about the truth value of the sentence as the information about the smaller set. We consider that the experimental design ruled out the informativity-based explanation.

Expectation-based account

Previous experiments (Urbach et al., 2015; Nieuwland, 2016) supporting the pragmatic account used participants' word knowledge to build specific expectations about the context. In contrast, in the current experiment, the contextual information was provided in a form independent of world knowledge, namely as pictures. Augurzky et al. (2020) argued that the predictability of both quantifiers was the same based on the contextual information. However, they did not fully reject the

expectation-based approach. Instead, they argued that the participants' ability to build expectations about the sentence was mediated by the complexity of the quantifiers. In other words, participants were more efficient in formulating predictions about *more than half* sentences than *fewer than half*. Because of the higher complexity of *fewer than half*, the expectations generation was delayed.

Our findings substantially extended the expectation-based interpretation of Augurzky et al. (2020). Firstly, the HsMM-MVPA analysis ruled out the interpretation that the generation of predictions for *fewer than half* was delayed due to the extra processing step. Moreover, the higher representational complexity of *fewer than half* was also not reflected in the extra stage. Secondly, our findings give additional insight into the interpretation of late positivity. We interpret late positivity as pragmatic P600. This interpretation is supported by the additional finding of P200 potential. We found that difference between quantifiers in P600 was preceded by the P200 difference (see the ERP analysis). Moreover, we found that the difference in P600 had a mostly parietal distribution. This pattern of results matches the previous findings on pragmatic P600 (Regel et al., 2014), which is usually preceded by P200.

Augurzky et al. (2020) provided two possible explanations of how the predictions could have been generated and to give rise to late positivity. According to the first explanation, participants encoded the picture in terms of the greater set (Clark, 1976). Based on the picture encoding, they could have immediately generated the expectations for a *more than half* sentence, but not *fewer than half*. The P600 potential can reflect the attempt to build expectations for *fewer than half* when it becomes clear that the negative quantifier has to be processed. The extra effort leads to the engagement of more cognitive resources and differences in EEG amplitude. This attempt is not fully successful as reflected by the lack of N400 difference on the adjective for *fewer than half* sentences. The second possibility was that participants did not have any expectations about the sentence after they saw the picture. They encoded both sets of shapes from the picture equally well. After the quantifier onset, participants tried to build expectations for both quantifiers. However, due to the higher complexity of *fewer than half*, participants struggled with building the expectations and were not successful after all. Augurzky et al. (2020) leaned toward the second explanation.

We suggest that the first explanation is more consistent with our findings. The representational complexity of *fewer than half* was not reflected in the HsMM-MVPA data. We failed to find evidence for the extra processing step. Neither did we find differences in stage durations after the quantifier onset, nor a delay in stage onsets for *fewer than half*.

The participants' expectations about the sentence continuation were reflected in the N400 effect after the adjective. We did not observe a delay in processing of *fewer than half*, but we hypothesized that the generation of the expectations may not have been successful. Our findings gave further insight into the interpretation of N400 in the Augurzky et al. (2020) study. While some studies link the N400

effect with a truth value evaluation (Augurzky et al., 2017, 2020), our results suggest that N400 reflects the ease of lexical retrieval of the adjective, but not yet the truth value evaluation. We associated the truth value evaluation with the later Stage 8. Our interpretation of the N400 effect is in line with the findings of Kounios and Holcomb (1992). Moreover, our analysis is consistent with the access/retrieval account for N400 supported by previous studies (Delogu et al., 2019). According to this account, N400 reflects lexical retrieval but not the integration process.

Our results give additional insight into the time course of processing quantified sentences. The first six stages of processing after the adjective onset are related to the validation of participants' expectations about the sentence continuation. Following this interpretation, we can give a meaningful explanation of the Augurzky et al. (2020) finding in the 450 to 800 ms time window after the adjective onset (also replicated in our ERPs analysis). The larger negativity for *more than half* true sentences preceded Stage 7, which we identified as the comparison stage. In this stage, participants compared the picture and sentence representations in their working memory. Stage 7 was longer for *fewer than half* than for *more than half* and for false sentence than true. We explain this difference in terms of working memory load during the comparison. Participants had to retain the information about the quantifier, the numerosity of objects, and the adjective. Assuming that they encoded the picture in terms of the bigger set if they verified the *more than half* true sentence, they only retained the matching information in their working memory. For the *more than half* false sentence, participants had to carry additional information that the set color and the adjective color did not match. While for *more than half* participants had to remember only the greater set, for *fewer than half* they had to retain information about both sets in their working memory. This explains the quantifier effect in Stage 7. Finally, for the *fewer than half* false sentence, participants also had to carry information about the mismatching color (similar to *more than half* false sentence). This explains the effect of the truth value.

We argue that participants evaluated the truth value of the sentence in Stage 8. This was accompanied by late negativity. This finding is consistent with evidence that late negativity, but not N400, reflects the truth value evaluation (Wiswede et al., 2013).

Additional support for our interpretation of Stages 7 and 8 comes from the analysis of long sentences (see Appendix C). We found a similar series of bumps from the first to the sixth bump in both sentences types. For both short and long sentences, participants compared two representations in working memory in Stage 7 before gaining the information about the further continuation of the sentence. We did not expect differences between the time course for the processing of short and long sentences until the evaluation of the truth value of the sentence in Stage 8. While the evaluation was necessary for the short sentences it was not needed for the long ones.

Consistent with this interpretation, we found that the topography of the seventh bump in long and short sentences differed. We observed a negative bump in short sentences and no such effect in long sentences. Furthermore, in Stage 9, we observed left frontal positivity in both types of sentences. In the long sentences, the positivity was also sustained in Stage 10. Left frontal positivity reflects the reaction inhibition (De Jong et al., 1990). Participants had to inhibit reactions in both sentence conditions: in the short sentences because they waited for a signal to respond, and in the long sentences because the sentences were to be continued. Therefore, we expected the bump to be related to the reaction inhibition in both sentence types. In contrast, the truth value evaluation was expected only in short sentences.

The final challenge for the expectation-based account is to explain the difference in stage durations in Stage 11. According to our interpretation, participants compared the picture and sentence representations in Stage 7 and evaluated the sentence's truth value in Stage 8. This would suggest that the last stage should not differ between conditions, as it only reflects the response execution. However, we found the effect of Quantifier on duration in Stage 11. This suggests that yet another factor affects the processing of negative quantifiers. Previous computational modeling studies (Schlotterbeck et al., 2020; Ramotowska et al., 2020a) found a difference in non-decision time between positive and negative quantifiers. Schlotterbeck et al. (2020) interpret this finding as an indication of the additional step in processing for negative quantifiers. Our analysis does not allow for direct mapping between Stage 11 and non-decision time, but it challenged the interpretation of Schlotterbeck et al. (2020).

Together, the differences in Stage 11 cannot be explained by either the expectation-based account or by the two-step model. As a possible explanation, we accept that the differences in the last stage were driven by the experimental procedure. We suggest validating our findings in a different experimental paradigm, e.g., without signal-to-response procedure and timeout on the responses (see Appendix C for timeout reaction time analysis). We noticed great variability in the duration of Stage 11 and reaction times. When the reaction times that exceeded the timeout were excluded from the analysis, the interaction effect was not significant (see Appendix C). Moreover, the analysis of Stage 11 with excluded timeout reaction times did not reveal differences between experimental conditions. This suggests that the differences in reaction times and Stage 11 were at least partially driven by the long responses. Therefore, we do not draw any firm conclusions and we leave the interpretation of Stage 11 for future work.

Future challenges

While our findings speak against the extra processing step for negative quantifiers, they can not rule out the two-step model hypothesis for other types of negatives. In particular, the two-step model originated from the studies on sen-

tential negation. It would be desirable to test for the extra processing step when the negation refers explicitly to the to-be-negated state of affairs. Once explicitly mentioned, the representation of the to-be-negated state of the affairs could be activated and processed in the extra stage. In conclusion, further studies should test the two-step model predictions in different experimental set-ups and with various types of negatives.

Moreover, while there is a general agreement that successful language comprehension requires building predictions about upcoming linguistic input (Grisoni et al., 2021, 2017), the explanation of mechanisms behind the expectations generation is less understood. The dynamic pragmatic account by Tian, Breheny, and Ferguson (2010) proposes such a mechanism by referring to the so-called Questions Under Discussion (QUDs). This account assumes that language users process the information that is already accommodated into the discourse context by the relevant QUDs faster. According to this approach, the positive questions are considered by the comprehenders as the default, because they are more frequent than negative questions. For example, when verifying the sentence "The glass is not empty," the comprehenders assume that the relevant QUD is "Is the glass empty?" rather than "Is the glass not empty?". By introducing the negative cleft structure, Tian et al. (2010) showed that negative sentences are not processed in the two-step manner. To directly test the dynamic pragmatic account, we would have to introduce the explicit manipulation of the QUDs that would affect the encoding of the picture. The exploration of how the manipulation of QUDs would affect the stage durations can be tested in future work.

5.4.4 Methodological implications

Our study has several methodological implications. Firstly, it showed the dissociation between behavioral measures (reaction times) and EEG-based measures (stage durations). Secondly, it also demonstrated the dissociation between different EEG measures. Specifically, we observed that the difference in EEG signal amplitude between quantifiers (in ERP analysis) was not reflected in the differences in stage durations.

The HsMM-MVPA method was previously applied to rather simple cognitive tasks such as associative recognition (Anderson et al., 2016; van Maanen et al., 2021), lexical decision tasks (Berberyan et al., 2021), or perceptual decision tasks (Berberyan et al., 2020). In contrast, we applied the HsMM-MVPA method to a complex linguistic task. This posed additional challenges to our analysis. For example, we could not analyze the processing of the whole sentence, but we had to select the constrained time windows. Moreover, the stages of processing we analyzed related to one stimulus overlapped with the display of new stimuli.

In the next sections of the discussion, we elaborate on these two key methodological aspects in more detail. We also discuss the limitations of our study.

The dissociation between different measures

We found different effects in the reaction time data and the duration of the stages. The stage durations did not reflect the interaction between Quantifier and Truth value in the behavioral data. Specifically, we tested the interaction effect in stages that predicted reaction times in all experimental conditions. While three stages contributed to the explanation of reaction times across conditions, some stages predicted only several conditions. The interaction in reaction times could be an effect of a unique combination of multiple stages, different for each condition. For example, Berberyan et al. (2020) found that two stages explained the non-decision time duration and three stages the decision time duration in a perceptual decision task. This finding suggests that the reaction times can be explained by the combined effect of multiple stages. The lesson to be learned from our analysis is that the mapping between the length of particular cognitive processes and reaction times might be equivocal. To draw firm conclusions, we have to jointly analyze behavioral and neural data.

We found a dissociation between behavioral and brain data, and also in different EEG measures. In the time window after the quantifier onset, we replicated the results of Augurzky et al. (2020). We also found greater late positivity in ERP analysis for *fewer than half* than for *more than half*. However, the greater neural activity for *fewer than half* was not reflected in stage duration differences between quantifiers.

The dissociation between the efficiency of cognitive processes and the neural response was also observed in the fMRI studies (see Kelly & O'Connell, 2015, for review) on decision-making. The steeper evidence accumulation signals the stronger evidence. When enough evidence is collected, the decision boundary is reached. Some studies linked the stronger BOLD response to faster evidence accumulation (e.g., Heekeren, Marrett, Bandettini, & Ungerleider, 2004), while others found the opposite effect (T. Liu & Pleskac, 2011). The link between BOLD response and evidence accumulation is mediated by the neural response after the decision boundary is reached. If the decision variable remains elevated after reaching the boundary, we can observe a positive correlation between the BOLD signal and the strength of evidence. If the decision variable drops, the opposite effect is observed.

The computational modeling experiments (Schlotterbeck et al., 2020) on negative and positive quantifiers showed that the steeper drift rate (faster accumulation) is associated with *more than half*. Following this finding, in the current study we would expect the evidence accumulation to be faster for *more than half* than *fewer than half*. We assume that participants gradually accumulate evidence together with the incremental processing of the sentence. Translating our results on the fMRI findings, we observe the negative relationship between evidence strength and neural response, meaning a lower EEG amplitude for *more than half* than *fewer than half*.

The confound of upcoming input

The analysis of incremental sentence processing with the presentation of words at fixed times inevitably leads to the confound when the processing stages of the first word overlap with the processing of the next word. This confound was also present in our analyses. For example, we made an assumption that in the 800 ms time window after the quantifier onset, we mostly measured the cognitive processes associated with quantifier processing. Nonetheless, we noticed that by the end of this time window, the next word was already present on the screen. The display of the new word could have impacted the consecutive stages of processing. We argue that the impact of the new word did not obscure our results. The next word displayed on the screen (German *als*) was highly predictable and the same in all experimental conditions. Moreover, it constituted part of the quantifier expression.

To obtain robust results and test the differences in processing stages that could be linked only to processing of the first word, we constrained the analysis to 500 ms after the quantifier onset (see Appendix C). In the shorter time window, only the first word was displayed on the screen. We ran two separate HsMM-MVPA models for each condition. We found that the 4-bump model fitted both quantifiers most accurately. The analysis again did not support the extra step of the processing hypothesis.

We compared the bumps is the first 500 ms of 800 ms time window with the new constrained time window. We noticed that in our first analysis, we had onsets of 5 bumps in the first 500 ms, while in the second was had only 4 bumps. We explain this difference by referring to the bump duration of 50 ms. The fifth bump in the 800 ms time window had an onset after 469 ms, which means that it was uncompleted in the first 500 ms. We observed that the bumps topographies were comparable in both time windows, while the bumps onsets were slightly different. Nonetheless, we conclude that we fully replicated the first five stages of processing in shorter (500 ms) and longer (800 ms) time windows.

Methodological limitations

We notice several methodological limitations of our study. Firstly, we found that some stages after the adjective onset had a different duration between conditions in the combined model. However, the mapped models that assumed these differences failed to outperform the combined model. One possible explanation of this finding is that the combined model was always better because it was fitted to the larger number of trials. In the current experiment, we had a relatively small number of trials per condition compared to previous HsMM-MVPA studies (e.g., Anderson et al., 2016; Berberyan et al., 2020; van Maanen et al., 2021). By combining the conditions in the combined model, we increased the number of trials from which the HsMM-MVPA algorithm estimated the bump magnitudes

and locations. For example, in the time window after the quantifier onset, we had a maximum of 160 trials per condition and participant, and in the time window after the adjective onset, we only had a maximum of 40 trials per condition and participant. The combined model was fitted on a maximum of 320 trials per participant after the quantifier onset and a maximum of 160 trials per participant after the adjective onset.

Secondly, we noticed that the greatest variability in model fit was present in the time window after the adjective onset for short sentences. The model fit results were more consistent for long sentences after the adjective onset (see Appendix C). This suggests that the small number of trials was not the only source of variability in model fit. The HsMM-MVPA model provided more stable results when the analyzed time windows were constrained to a time limit instead of by participants' responses. Participants' responses may have introduced greater variability into the EEG data. One source of variability may be the movement artifacts, even though we cleaned the data and removed trials containing larger artifacts. The variability could also be a result of the experimental procedure. For example, while participants waited for a signal to the response, their attention may have drifted away from a task. Moreover, the timeout procedure may have introduced pressure on participants' responses. Augurzky et al. (2020) argued against the speed-accuracy trade-off. Nonetheless, the deadlines for the response affect the decision-making process, as reflected in the modeling data (Katsimpokis et al., 2020). The response threshold declines under deadline manipulation.

Finally, the variation in data was also present in reaction times (see Appendix C). The timeout reaction times changed the typical reaction time distribution, and the log-transformation failed to compensate for it. The timeout trials constituted a large proportion of trials. Moreover, we noticed that they were more frequent for *fewer than half* than for *more than half*. The tendency for participants to exceed the time limit more for one quantifier could be a result of processing difficulties associated with it.

The reaction times that exceeded the time limit are challenging to interpret because we do not know their source. On the one hand, if the timeout reaction times are due to participants' attention drifting away before responding, it would be better to exclude trials with a timeout. The focus of attention can lead to different processing stages (van Maanen et al., 2021). However, we would have to exclude the timeout trials from all of our analyses because we do not know at what timepoint participants were distracted. This would lead to a significant reduction in the number of trials. On the other hand, the timeout reaction times were associated more with *fewer than half* than *more than half*. They could also be a relevant source of information about processing differences. From a methodological perspective, the decision to exclude timeout trials is unequivocal.

5.4.5 Conclusions

In this study, we challenged the two-step processing hypothesis in quantified sentences. By using a novel method to analyze the EEG data, we estimated the number of processing stages for sentences with two quantifiers *more than half* and *fewer than half*. To the best of our knowledge, this is the first study that directly tested the difference in processing stages in quantified sentences and directly addressed the two-step processing hypothesis. We provided an interpretation of the processing stages and linked them to the predictions of the expectation-based account. We also indicated future challenges to the pragmatic account that could not explain of all our results.

Chapter 6

Does ease of learning explain quantifier universals?[1]

Abstract All natural languages share common properties called universals. In a domain of quantification three semantic universals were discovered: monotonicity (convexity), quantity, and conservativity. Researchers have been trying to explain the origin of semantic universals for decades. In this study, we tested one of the proposed explanations, namely the learnability hypothesis. According to this hypothesis, quantifiers that satisfy universals are easier to learn and therefore more likely to be lexicalized in natural language. We tested the learnability hypothesis in a large-scale online between-subjects design experiment, in which participants learned a new quantifier *gleeb*. *Gleeb* corresponded to one the following quantifiers: monotone and quantitative *at least 3* and *at most 2*, non-monotone *between 3 and 6*, non-convex *at most 2 or at least 7*, non-quantitative *the first 3*, *the last 3*, conservative *not all*, and non-conservative *not only*. We found that monotonicity universal had a strong effect on the speed of acquisition: participants learned monotone quantifiers faster than non-monotone quantifiers. In contrast, conservativity did not have any effect on learning. Moreover, participants had higher accuracy for quantitative quantifiers than non-quantitative ones at the beginning of the experiment. In conclusion, learnability could be considered as one of the pressures in shaping the semantic universals, but some universals seek different explanations.

6.1 Introduction

Natural languages share common properties. For example, the phonology level holds the universal "all languages have consonants and vowels" (Hyman, 2008). At the lexicon level, it is true that "every human language has proper names"

[1]This chapter is based on the manuscript: Ramotowska, van Maanen, and Szymanik (2022). Does ease of learning explain quantifier universals? (unpublished manuscript).

(Universal 1321, Hockett, 1963). In the number system domain, Greenberg (1978) proposed the universal "in every numerical system some numbers are expressed by basic terms" (Universal 528). Explaining the origins of universal properties is one of the main goals of linguistic theory. Famously universal properties have been postulated in the domain of quantification. In this paper, we focus on what have historically been sees as the most prominent semantic universals: monotonicity (convexity), quantity, and conservativity (Barwise & Cooper, 1981; Keenan & Stavi, 1986; Peters & Westerståhl, 2008).

In natural language, quantifiers are logical words that express quantities, for example: *most*, *at least 6*, *all*, *at most 10*, *between 11 and 19*, etc. Quantifiers have well-defined mathematical properties and therefore constitute a good case study example.

Some quantifiers are lexicalized in natural languages. For example, in English, *most*, *all*, *some*, *few*, and *none* are lexicalized, but the quantifier that expresses a quantity *between 3 and 10* is not lexicalized. In contrast, in Polish, this quantifier is lexicalized as *kilka*. It seems that lexicalization plays an important role in natural language. It is commonly assumed that languages lexicalized the most crucial concepts to make communication more efficient. To have concepts expressed by just a single word speeds up communication and reduces the memory load.

According to the universal hypothesis, lexicalized quantifiers across languages are convex, quantitative, and conservative (Barwise & Cooper, 1981; Keenan & Stavi, 1986; Peters & Westerståhl, 2008). While this hypothesis is well-argued theoretically, the origin of semantic universals in the domain of quantifiers remains a matter of debate. One of the main questions of this debate is: why do universals exist? Researchers have been trying to explain the emergence of semantic universals by different pressures. One of them concerns the constraints on the human ability to learn concepts, i.e. learnability. According to this approach, universals exist because they allow for efficient meaning acquisition (Steinert-Threlkeld & Szymanik, 2019; Carcassi, Steinert-Threlkeld, & Szymanik, 2019). It has been hypothesized that *quantifiers satisfying universals should be easier to learn than quantifiers that do not satisfy universals* (see examples of this hypothesis in: Hunter & Lidz, 2013; Steinert-Threlkeld & Szymanik, 2019; Carcassi, Steinert-Threlkeld, & Szymanik, 2019). We will call this claim a learnability hypothesis.

The ease of learning can be understood in two ways. Firstly, universals constrain the space of meanings used in natural language, and learning from a smaller space is easier than from larger space (e.g., Barwise & Cooper, 1981; Keenan & Stavi, 1986). However, Steinert-Threlkeld and Szymanik (2019) argued that this form of learnability hypothesis does not explain which constraints are universal, because any type of constraint reduces the space of meaning. Secondly, the meanings that satisfy universals are themselves easier to acquire (Steinert-Threlkeld & Szymanik, 2019). Steinert-Threlkeld and Szymanik (2019) proposed a plausible bridging assumption between meaning acquisition and meaning structure

in natural languages: the meanings that are easier to learn are more likely to be lexicalized. Humans tend to express easy-to-learn meanings in their lexicon. In this way, the pressure of learnability explains the link between semantic universals and the meaning structure in languages. The quantifiers that satisfy the semantic universals are easier to learn and, as a consequence, their meanings are lexicalized. The universal claim is that *lexicalized quantifiers in natural languages satisfy the semantic universals*[2].

The learnability hypothesis generates testable predictions. Humans should learn the meanings that satisfy the universal properties faster. However, previous research has provided mixed experimental evidence concerning the link between learnability and semantic universals. In this study, we tested the learnability hypothesis with respect to three semantic universals of quantifiers: monotonicity (convexity), quantity, and conservativity. We ran a large-scale online artificial learning experiment with adult participants. In this experiment, participants learned a new word *gleeb* which corresponded to one of eight quantifiers: *at least 3*, *at most 2*, *the first 3*, *the last 3*, *between 3 and 6*, *at most 2 or at least 7*, *not all*, and *not only*. In the next sections, we explain the tested semantic universals and summarize experimental and modeling studies on semantic universals.

6.1.1 Quantifiers

Quantifiers[3] denote the relationship between two sets: set A, called restrictor, and set B, called scope. In natural language, quantifiers correspond to determiners (Det), which in combination with common nouns create the noun phrase (NP), e.g., *most dogs*, *some cats*, *few students*. When combined with a verb phrase (VP) they constitute quantified sentences e.g., *most dogs chase cats*, *some cats like fish*, *few students are brilliant*. We write $< M, [\![NP]\!], [\![VP]\!] > \in Det$ whenever the quantified sentence is true in the model, where M stands for a domain. We will use a generalized quantifier framework to formally define quantifiers and universals. We will use notation $< M, A, B > \in Q$, meaning that the sets A, B, and domain M belong to quantifier Q. We define the quantifiers as a relationship between sets A and B. Below, we present the formal definition for a few quantifiers commonly used in natural language:

$$[\![some]\!] = \{< M, A, B >: |A \cap B| \geq 1\}$$
$$[\![all]\!] = \{< M, A, B >: A \subseteq B\}$$

[2]The universal claim is sometimes formulated as: *simple quantifiers in natural language satisfy the semantic universals*. The notion of simplicity is vague. For example, *most* is considered a lexicalized quantifier that satisfies universals, yet it is a complex quantifier according to Hackl's (2009) analysis (*most* means *many* + est). In this paper, we will use the notion lexicalized to avoid confusion.

[3]In this paper, we focus only on quantifiers of type $< 1, 1 >$, see Peters and Westerståhl (2008).

$[\![most]\!] = \{< M, A, B >: |A \cap B| > |A|/2\}$

6.1.2 Monotonicity and convexity (connectedness)

Monotonicity is a prominent semantic universal which refers to the entailment pattern. The quantifier is upward monotone (in its right argument) if it does not change the truth value with an increase in its scope. The quantifier is downward monotone if its truth value is preserved while its scope decreases. Consider these intuitive examples:

(1) *At least 6* students passed the difficult exam.
(2) *At least 6* students passed the exam.
(3) *At most 6* students passed the difficult exam.
(4) *At most 6* students passed the exam.
(5) *Between 3 and 6* students passed the difficult exam.
(6) *Between 3 and 6*students passed the exam.

The Sentence (1) entails Sentence (2). This is because the quantifier *at least 6* is upward monotone and by changing its scope to more general (the term "exam" is more general than the term "difficult exam"), the truth value of the sentence is preserved. In contrast, Sentence (4) entails Sentence (3), because the quantifier *at most 6* is downward monotone. The sentence with a downward entailing quantifier has the same truth value with the decreasing scope from "exam" to "difficult exam". The quantifier like *between 3 and 6* is not monotone: neither Sentence (5) does not entail (6) nor does (6) entail (5).

We call a quantifier monotone if and only if it is either upward monotone or downward monotone. Formally, monotonicity is defined as in (7) and (8) (Barwise & Cooper, 1981)[4]:

(7) The quantifier Q is upward monotone iff if $< M, A, B > \in Q$ and $B \subseteq B'$, then $< M, A, B' > \in Q$.
(8) The quantifier is downward monotone iff if $< M, A, B > \in Q$ and $B' \subseteq B$, then $< M, A, B' > \in Q$.

Convexity or connectedness is a weaker version of monotonicity[5]. A monotone quantifier is also convex but not vice versa. Informally, this property indicates that if two objects a and c have a property X, then any object b that is in between has this property. For example, "*Between 4 and 10* triangles are red" is convex, because triangles' cardinalities in the range four to ten have the property of being red. A convex quantifier can be defined in terms of conjunction of

[4]See also Steinert-Threlkeld and Szymanik (2019), p. 6.
[5]Peters and Westerståhl (2008) use the term *continuous quantifiers* for convex quantifiers (p. 168).

monotone increasing and monotone decreasing quantifiers, for example, *between 4 and 10* means *at least 4* and *at most 10*. Formally, convexity for quantifiers can be defined as in (9):

(9) The quantifier Q is convex iff if $< M, A, B' > \in Q$, $< M, A, B'' > \in Q$ and $B' \subseteq B \subseteq B''$ then $< M, A, B > \in Q$.

Following Barwise and Cooper (1981), we defined a monotonicity universal (MU) as:

MU: All lexicalized determiners express monotone quantifiers or conjunctions thereof.

According to MU, the lexicalized quantifiers in natural languages are monotone or convex. The universal claim, however, does not mean that the same quantifiers have to be lexicalized in all natural languages (recall the example of the convex quantifier *kilka* in Polish and *between 3 and 10* in English).

Partial empirical evidence for a role of monotonicity and convexity universals in learnability of quantifiers comes from experiments on adult humans (Chemla, Buccola, & Dautriche, 2019), and an artificial model of learning, e.g., neural networks (Steinert-Threlkeld & Szymanik, 2019). In the former experiment, participants learned three types of rules: monotone (e.g., "There are 3, 4, or 5 red circles."), connected (e.g., "There are 2, 3, or 4 red circles.") or non-connected (e.g., "There are 1, 2, or 4 red circles."). Participants saw a display with five circles in different colors and were asked to assess whether the display was consistent with a rule. They had to infer the rule based on feedback in each trial. Chemla, Buccola, and Dautriche (2019) showed that participants learned the non-connected rule more slowly than the monotone rule.

Moreover, Steinert-Threlkeld and Szymanik (2019) showed the monotonicity role in learnability by using a long short-term memory recurrent neural network model of learning. The network learned faster both upward (*at least 4*) and downward (*at most 3*) monotone quantifiers than non-convex quantifier (*at least 6 or at most 2*).

Finally, additional evidence for the role of the monotonicity universal comes from the language evolution experiments. Carcassi, Steinert-Threlkeld, and Szymanik (2019) showed the emergence of monotone quantifiers in their iterative learning experiment with neural network agents. Moreover, monotonicity universal plays an important role in other language domains, for example, scalar adjectives (Carcassi, Schouwstra, & Kirby, 2019). In the iterated language experiment, Carcassi, Schouwstra, and Kirby (2019) showed that learnability is one of the pressures for monotonicity to evolve. They fitted three computational models in the Rational Speech Act framework (Frank & Goodman, 2012) and showed that the model with combined learnability pressure and pragmatic skills

of agents led to the evolution of monotone adjectives.

6.1.3 Quantity

Quantity,[6] often known in the literature as isomorphism invariance (Peters & Westerståhl, 2008), is traditionally a part of the definition of generalized quantifiers (Mostowski, 1957). Intuitively, quantity means that the truth value of a quantifier sentence depends only on the number of elements satisfying relevant properties and not on the order or nature of the elements. In particular, the manner of presentation of the elements should be irrelevant. Therefore, quantifiers such as *the first 3* or *the last 3* do not satisfy the quantity universal. Formally, quantity is defined as isomorphism invariance (Peters & Westerståhl, 2008)[7]:

(10) The quantifier Q is isomorphism-invariant iff if $< M, A, B > \cong < M', A', B' >$, then $< M, A, B > \in Q$ if and only if $< M', A', B' > \in Q$.

The quantity universal (QU) says that (Peters & Westerståhl, 2008):

QU: All lexicalized determiners are isomorphism-invariant.

As far as we know, the effect of quantity on learnability was not tested in human participants, but the artificial learning model learned the quantitative quantifier *at least 3* faster than the two non-quantitative quantifiers *the first 3* and *the last 3* (Steinert-Threlkeld & Szymanik, 2019).

6.1.4 Conservativity

Conservativity may be the most famous among quantifier universals (Peters & Westerståhl, 2008). Intuitively, conservativity means that the quantifier restrictor restricts the sentence topic. For example, the sentence "*Not all* professors attended the conference." contains a conservative quantifier *not all*. The sentence says that the set of professors is not included in the set of people who attended a conference, and the sentence focuses only on the set of all professors and the professors who attended the conference. The conservativity of *not all* can be tested in a linguistic test. Consider the examples below:

(11) *Not all* professors attended the conference.
(12) *Not all* professors are professors who attended the conference.

Sentences (11) and (12) are equivalent. In contrast, the sentence "*Not only* professors attended the conference" contains a non-conservative quantifier *not*

[6]The term "quantity" was introduced by van Benthem (1986).
[7]See also Steinert-Threlkeld and Szymanik (2019), p. 7.

only. The sentence refers to a set of people who attended the conference, and who are not professors, for example students. Formally, conservativity is defined as[8]:

(13) The quantifier Q is conservative iff: $< M, A, B >\in Q$ iif $< M, A, A\cap B >\in Q$.

The conservativity universal (CU) claims that (Barwise & Cooper, 1981; Keenan & Stavi, 1986; see Peters & Westerståhl, 2008 for discussion)[9]:

CU: All determiners are conservative.

The role of learnability in the explanation of conservativity as a semantic universal is the most controversial. Some studies have shown that conservativity facilitates learning in children (Hunter & Lidz, 2013), while others have not replicated this finding (Spenader & de Villiers, 2019). Hunter and Lidz (2013) showed that children learned the artificial quantifier *gleeb* faster when its meaning corresponded to the conservative quantifier *not all* than to the non-conservative quantifier *not only*. Spenader and de Villiers (2019) failed to replicate Hunter and Lidz's (2013) finding both in adults and children. Neither groups was successful in learning the new quantifier. Importantly, they were equally unsuccessful regardless of whether the quantifier satisfied or not the conservativity universal or not. Moreover, Spenader and de Villiers (2019) applied a different experimental paradigm, a situation verification with correction, to further test the effect of conservativity on learnability. They found that children learned the conservative quantifier *all* and the non-conservative quantifier *only* equally well, but did not learn the conservative quantifier *not all* and the non-conservative quantifier *not only*. Taken together, all four experiments failed to show the difference between conservative and non-conservative quantifiers. It is worth mentioning that Hunter and Lidz's (2013) experiment consisted of training and test blocks, both very short with only five trials. The Spenader and de Villiers's (2019) novel paradigm, in turn, consisted of 10 training and 10 testing trials. The experiment was, therefore, also fairly short and, as admitted by the authors, the lack of difference between quantifiers *not all* and *not only* could be a result of not enough training trials.

Following the linguistic literature, Steinert-Threlkeld and Szymanik (2019) predicted that monotonicity and quantity should positively affect the ease of quantifier learning, while conservativity should be explained independently from learning simplicity. The neural networks learned conservative and non-conservative

[8]See also Steinert-Threlkeld and Szymanik (2019), p. 8.

[9]Barwise and Cooper (1981) did not use the term "conservativity". They formulated the determiner universal (Universal 3) on p. 179. The term "conservativity universal" was firstly used by Keenan and Stavi (1986).

quantifiers at the same rate. This result was consistent for a pair of quantifiers tested on children, namely *not all* and *not only*, as well as, for the pair *most* and artificial non-conservative quantifier M (meaning of M was $|A| > |B|$).

6.1.5 Current experiment - predictions

The goal of this study was to test the learnability hypothesis for multiple semantic universals in a large-scale experiment. Following Steinert-Threlkeld and Szymanik (2019) and Chemla, Buccola, and Dautriche (2019), we predicted under the learnability hypothesis that quantifiers that satisfy monotonicity, convexity, and quantity universals will be easier to learn. Based on Steinert-Threlkeld and Szymanik (2019) and Spenader and de Villiers (2019), we did not expect to find the learnability effect for conservativity.

We adopted a word-learning experimental paradigm. In our experiment, participants learned the meaning of a new, lexicalized quantifier, for example, *gleeb*, by experiencing the conditions that satisfy and do not satisfy *gleeb*. We included two monotone quantifiers *at least 3* (upward) and *at most 2* (downward) as well as non-monotone quantifiers *between 3 and 6* and *at most 2 or at least 7*. We tested the convex quantifier *between 3 and 6* versus the non-convex quantifier *at most 2 or at least 7*. For quantity, we compared the quantitative quantifiers *at least 3* and *at most 2* with the non-quantitative quantifiers *the first 3* and *the last 3*. For conservativity, we chose *not all* (conservative) versus *not only* (non-conservative). All of the chosen quantifiers are complex quantifier expressions in English because they are not lexicalized.

The choice of quantifiers was motivated by the previous studies. These studies (Hunter & Lidz, 2013; Spenader & de Villiers, 2019; Steinert-Threlkeld & Szymanik, 2019; Chemla, Buccola, & Dautriche, 2019) applied the minimal pair methodology. The idea behind this methodology is to choose the pairs of quantifiers that differ in universal properties but are otherwise comparable, and test them against each other. Although this methodology has several limitations, it is the most feasible paradigm to test with human subjects.

We chose numerical monotone, convex, and non-convex quantifiers for comparison with Chemla, Buccola, and Dautriche (2019). Although participants learned rules rather than quantifiers in their experiment, these rules could be translated into quantifiers like *at least x*, *at most y*, *between x and y*, and *at most x or at least y*. In our study, the non-convex quantifier was a negation of the convex quantifier.

We chose the same non-quantitative quantifiers as Steinert-Threlkeld and Szymanik (2019). The non-quantitative quantifiers in their study referred to numerical information and the order of presentation. For example, the sentence "*The first 3* triangles are red" requires a presentation of *at least 3* red triangles, as well as ordering. There have to be three red triangles at the beginning to satisfy the sentence. Therefore, we contrasted the non-quantitative quantifiers with

monotone quantifiers[10].

We chose the same conservative vs. non-conservative pair of quantifiers as Hunter and Lidz (2013), Spenader and de Villiers (2019), and Steinert-Threlkeld and Szymanik (2019). We selected this pair, because it is comparable in terms of complexity (van de Pol et al., 2019, 2021) and because both quantifiers are present in natural language. We did not include the pair *most* and *M* in our study, because *most* is already lexicalized in English, while *M* is an artificial quantifier. Nonetheless, Steinert-Threlkeld and Szymanik (2019) showed that the results for the *not all* and *not only* pair are qualitatively identical to the results for *most* and *M*.

6.2 Methods

6.2.1 Participants

We recruited 246 participants for our experiment. We excluded only the incomplete data of 7 participants. The final sample consisted of 239 participants: *at least 3* ($N = 30$), *at most 2* ($N = 30$), *between 3 and 6* ($N = 31$), *at most 2 or at least 7* ($N = 30$), *the first 3* ($N = 30$), *the last 3* ($N = 30$), *not all* ($N = 29$), *not only* (N = 29). The mean age in the final sample was 36 years (range 18 – 70). All participants were native speakers of English. The experimental procedure was accepted by the Ethics Committee of the Faculty of Social and Behavioral Sciences of the University of Amsterdam, and all participants gave informed consent.

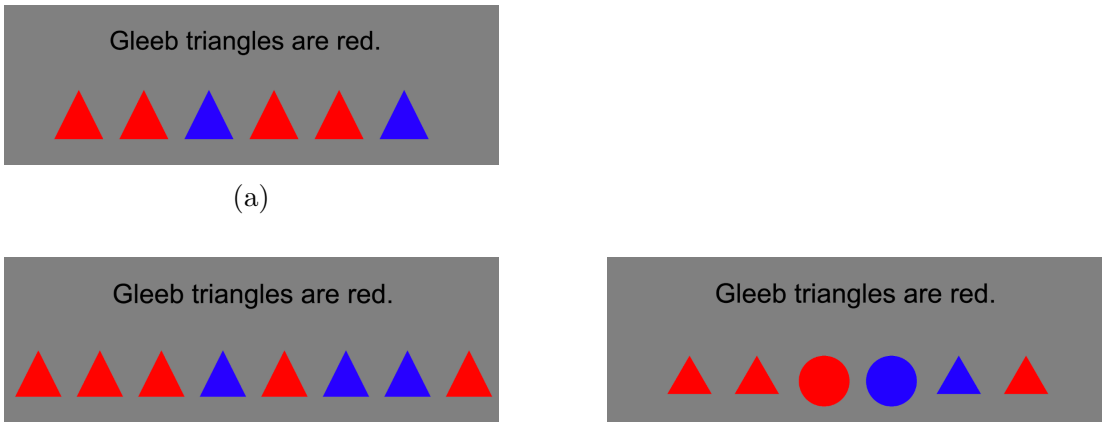
6.2.2 Materials and design

The experiment had a between-subjects design, meaning that each participant learned only one quantifier. Each trial in the experiment consisted of two displays. On the first screen, the participants saw a sentence "Gleeb triangles are red." and between one and eight geometric objects positioned in the row below the sentence (see Figure 6.1 for example stimuli). The objects were displayed in a row because their order was important for non-quantitative quantifiers. The objects could be either blue or red.[11] There was always at least one target – a red triangle – on the screen. In the next sections, we describe the stimuli in more detail for each quantifier. Participants had to press the T key if they thought that the sentence was true or the F key if they thought that the sentence was false.

[10]Note that all monotone (*at least 3* and *at most 2*) and non-quantitative (*the first 3* and *the last 3*) quantifiers are conservative and that non-quantitative quantifiers are monotone. For example, from the sentence "*The first 3* triangles are light red." it follows that "*The first 3* triangles are red." Therefore, *the first 3* is monotone increasing.

[11]Red and blue are a good pair because blue-red color blindness is found only in a rare complete color blindness, see National Eye Institute (`https://www.nei.nih.gov/`).

On the second screen, participants saw the same geometric objects (in the same colors and position on the screen) and feedback about their answers. They saw feedback "Correct!" displayed in green color if their answer was correct and "Wrong!" displayed in red color if their answer was incorrect. The feedback for correct responses was displayed for 1 second and the feedback for an incorrect response for 3 seconds to encourage participants in providing correct answers. After the feedback disappeared the new trials started. The whole experiment consisted of 96 trials, 8 implicit blocks for 12 trials. We scaled the length of the experiment based on findings by Chemla, Buccola, and Dautriche (2019). We decided that 96 trials should be enough for participants to learn even more difficult quantifiers, while keeping the experiment short enough to avoid a high drop-out rate (Schnoebelen & Kuperman, 2010). There was an equal number of true and false trials in each block. We did not include breaks between blocks. The block procedure gave us more control over the randomization procedure.



(a)

(b)

(c)

Figure 6.1: Example stimuli used in the experiment. 6.1a: Example stimuli type for monotone, convex, and non-convex quantifiers. Six triangles (four red) in the random order. For this stimulus, participants should have provided a "true" response for quantifiers: *at least 3* and *between 3 and 6*, and "false" for *at most 2* and *at most 2 or at least 7*. 6.1b: Example stimuli type for quantitative vs. non-quantitative quantifiers, eight triangles (five red), three first triangles are red. For this stimulus, participants should have provided a "true" response for quantifiers: *the first 3* and *at least 3*, and "false" for *at most 2* and *the last 3*. 6.1c: Example stimuli type for conservative vs. non-conservative quantifiers, six figures (four triangles, two circles). For this stimulus participants should have provided a response "true" for quantifiers *not all* and *not only*.

Stimuli and randomization procedure for monotonicity and convexity

For quantifiers *at least 3*, *at most 2*, *between 3 and 6*, and *at most 2 or at least 7* we only used one type of figure – triangles. We generated the list of all possible combinations of trials given that the number of triangles varied between one and eight and there was always at least one red triangle. Then we assigned the correct response to each trial based on the quantifier's meaning. Because the number of possible trials per numerosity of red triangles was not equal (e.g., there was only one possible trial when eight triangles were red and eight possible trials with one red triangle), we had to apply undersampling and oversampling procedures. In each block, we randomly drew six trials where the quantifier was true and six where it was false in such a way that all numerosities of the red triangle were represented. In this way, we made sure that participants saw each numerosity at least eight times in the experiment (in each of eight blocks). We believe that we achieved the optimal trade-off between variability in the stimuli and balance between numerosities of red triangles, which could affect the learning of quantifiers. The position of the triangles was randomized in each trial and centered in the middle of the screen.

Stimuli and randomization procedure for quantity

As for monotonicity and convexity, we also generated all possible trials for quantifiers: *the first 3* and *the last 3*. We used only red and blue triangles and there was always one red triangle on the screen. For consistency with the monotonicity and convexity experimental design, we also used a block design for non-quantitative quantifiers. In each block, six trials with correct answer "true" and six with correct answer "false" were randomly selected. We randomized the position of the first and last three triangles separately. To increase variation in stimuli, all numerosities of the target triangles were displayed in each block.

Stimuli and randomization procedure for conservativity

For conservativity, we used two types of figures to generate stimuli for the non-conservative quantifier *not only*. The sentence "*Not only* triangles are red" is true only if there is also another type of red figure. We therefore chose to have four triangles and four circles. One figure of each type was always displayed, and there was always one red triangle on the screen (circles could all be blue). The block procedure was the same as for other quantifiers for a consistency purposes. We randomized the position of all figures in each trial. To increase variation in stimuli, all numerosities of the target triangles (or circles for *not only*) were displayed in each block.

6.2.3 Procedure

The experiment was implemented in PsychoPy version 2020.2.4. and PsychoJS version 2020.1 and stored on the Pavlovia platform (`https://pavlovia.org/`). Participants were recruited via the Prolific platform (`https://www.prolific.co/`) and provided with the link to the experiment. After they filled in basic demographic information (gender and age) and gave informed consent for participation, they read the instructions for the experiment and proceeded with the task. After they completed the experiment, they were asked about the meaning of *gleeb*.

6.2.4 Statistical analysis

Firstly, we wanted to generally assess participants' performance in the task. Specifically, we checked how many participants learned each of the quantifiers. We assumed that participants learned the quantifier if their performance was above chance level. To have statistically significant evidence that participants performed above chance, they needed to provide at least 10 out of 12 correct responses in the final block (binomial test *binom.test* in R, $p = 0.04$). This means that the learning criterion was that the accuracy threshold equaled 0.83 in the final block.

We used the *stan.glmer* function from the *rstanarm* R package (Stan Development Team, 2020) to analyze data using the Bayesian mixed-effects logistic regression model with a linking function logit. We defined the universals as the following contrasts: monotonicity (*at least 3*, *at most 2* vs. *between 3 and 6*, *at most 2 or at least 7*), convexity (*between 3 and 6* vs. *at most 2 or at least 7*), quantity (*at least 3*, *at most 2* vs. *the first 3*, *the last 3*) and conservativity (*not all* vs. *not only*). We ran a model for each universal. We included Universal (0 not satisfying universal, 1 satisfying universal), Quantile of trials (called Training, 0 the first 25% of trials and 1 the last 25% of trials) and their interaction to predict participants responses (1 correct and 0 incorrect). We included the by-subject random intercept into all models.

All priors were set to default. The prior of each intercept was a normal distribution with a mean of 0 and a standard deviation of 10. The priors of the coefficients priors had a mean of 0 and a standard deviation of 2.5. To estimate the posterior distribution, we ran Hamiltonian Markov Chain Monte Carlo (MCMC) sampling with four chains. Each chain contained 1,000 warm-up samples and 1,000 samples. For a model diagnostic, we looked at *Rhat* values, a number of effective samples, autocorrelations in chains, and posterior predictive checks. We assessed the models using the *ShinyStan* Version 2.5.0 (Stan Development Team, 2017).

6.2.5 Frequency analysis of quantifiers

In our experiment, we tested adult native English speakers had who already acquired the language. The frequency of quantifiers of choice could affect the ease of learning. To control for this confound, we measured the frequencies of quantifiers included in minimal pairs based on the Corpus of Contemporary American English (COCA) (Davies, 2008-). Since the update in March 2020, the corpus contains more than one billion words (precisely 1,001,610,938). The corpus includes words from various sources of spoken language, movies and television subtitles, webpages, blogs, academic texts, newspapers, popular magazines, and fiction.

We measured the frequency of the exact quantifiers that we taught our participants as well as similar quantifier constructions. For example, we measured the frequency of *at least 3* (together with other forms of this quantifier such as *more than 2*), and the constructions such as *at least NUMBER NOUN*. For numerical quantifiers, we searched the number expressed by the word (*at least three*) or by the Arabic number symbol (*at least 3*). The frequency analysis is included in Appendix D.

6.3 Results

Firstly, we checked how many participants in the final block performed above chance by assessing the accuracy threshold of 0.83, meaning that they learned the quantifier (Figure 6.2). The results show that overall our task was quite difficult and, for most of the quantifiers, only around two-thirds of participants performed significantly above chance. The most difficult quantifier appeared to be *at most 2 or at least 7* as the lowest number of participants performed above chance on this quantifier.

Next, we tested whether participants learned the quantifiers that satisfied the semantic universals faster. We compared the mean accuracies of all participants in the first 25% of the trials and the last 25% of the trials (Figure 6.3). We expected higher accuracy in the last 25% of the trials and a larger training effect for quantifiers that satisfied semantic universals. The effect of Universal can be seen by comparing the position of the circles and triangles between quantifiers and universals, and the effect of Training during the experiment can be observed by comparing the length of the black arrows.

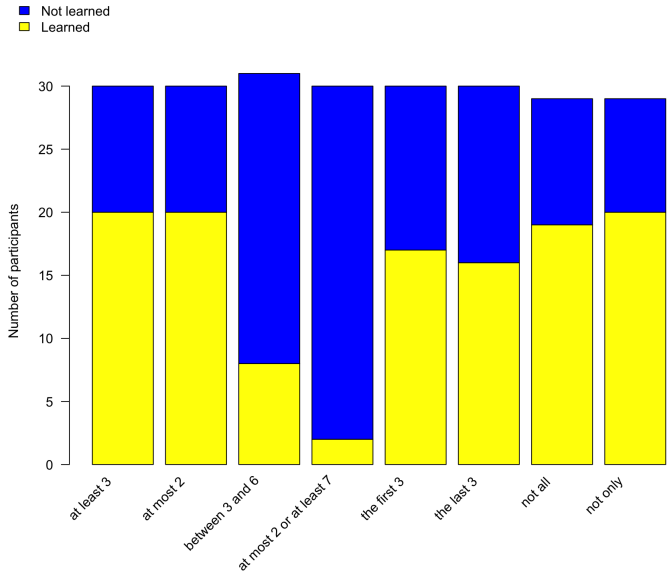
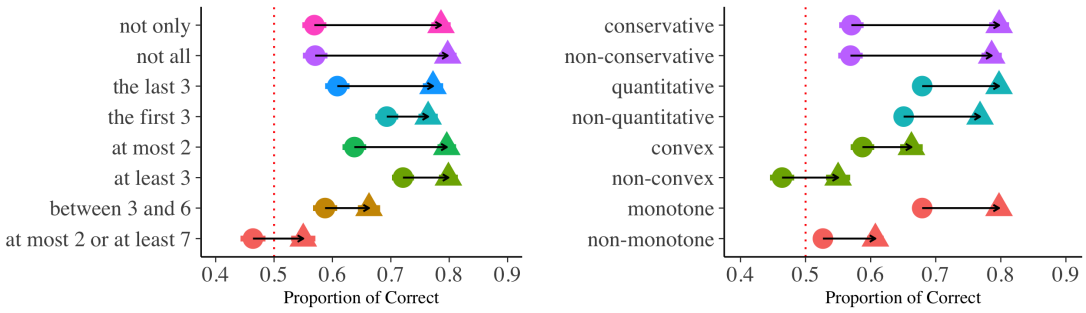


Figure 6.2: Number of participants who did (yellow) or did not (blue) achieve the 83% learning criterion per quantifier.



(a) (b)

Figure 6.3: The proportion of correct responses per quantifier (6.3a) and per universal (6.3b) for the first 25% of trials (circles) and the last 25% of trials (triangles). The color bars indicate the standard error of mean. The dashed red line indicates 50% accuracy. The arrows indicate the increase in accuracy between the first and the last 25% of trials.

Figure 6.3a presents the summary of the mean accuracy of participants in the first 25% of the trials and the last 25% of the trials for each quantifier. It shows a firm between-quantifier variability in participants' performance already occurring in the first 25% of trials. In the last 25% of trials, participants achieved different accuracy depending on the quantifier. The highest mean accuracy in the last 25% of trials was around 80%, and the lowest around 55%.

Figure 6.3b summarizes mean accuracy in the first 25% of the trials and the last 25% of the trials broken down into universals. It shows large differences in accuracy between monotone and non-monotone quantifiers and convex and non-convex quantifiers. The difference between conservative and non-conservative quantifiers is the smallest.

To test these effects statistically, we fitted a Bayesian mixed-effects logistic regression model. All Bayesian logistic regression models satisfied the assumptions of the test (see Table D.1 in Appendix D for more details). We reported the mean regression coefficients with 95% credible intervals (CI). Crucially, we tested whether the CI of each coefficient did not include the zero value. The posterior distributions of the effects with 95% credible intervals are presented graphically in Figures D.1 to D.3 in Appendix D.

6.3.1 Monotonicity

For the monotonicity universal, the mean intercept was $\beta = 0.12$ with the credible interval [CI: -0.06, 0.3] including zero, the Universal effect was $\beta = 0.69$ [CI: 0.4,3 0.95], the Training effect was $\beta = 0.35$ [CI: 0.2, 0.5], and the interaction effect was $\beta = 0.32$ [CI: 0.09, 0.56]. We computed for Universal effect $\Pr(\beta > 0) = 1$, for Training effect $\Pr(\beta > 0) = 1$, and for interaction effect $\Pr(\beta > 0) = 0.996$.

Additionally, we tested the model with pairs of quantifiers: *at least 3* vs. *between 3 and 6* and *at most 2* vs. *between 3 and 6* to zoom in on differences between quantifiers. For the contrast between the upward monotone quantifier and non-monotone quantifier, we found that the mean intercept was $\beta = 0.37$ [CI: 0.12, 0.63], the Universal effect was $\beta = 0.63$ [CI: 0.27, 0.99], the Training effect was $\beta = 0.34$ [CI: 0.13, 0.55], and the interaction effect was $\beta = 0.12$ with the credible interval [CI: -0.22, 0.46] including zero. The Universal effect had $\Pr(\beta > 0) = 0.99975$, the Training effect $\Pr(\beta > 0) = 0.99925$, and the interaction effect $\Pr(\beta > 0) = 0.76$.

For the contrast between the downward monotone quantifier and non-monotone quantifier, we found that the mean intercept was $\beta = 0.38$ [CI: 0.11, 0.65], the Universal effect was $\beta = 0.24$ with the credible interval [CI: -0.13, 0.62] including zero, the Training effect was $\beta = 0.34$ [CI: 0.12, 0.56], and the interaction effect was $\beta = 0.52$ [0.19, 0.86]. The Universal effect had $\Pr(\beta > 0) = 0.89$, the Training effect $\Pr(\beta > 0) = 0.99925$, and the interaction effect $\Pr(\beta > 0) = 0.99925$.

Taken together, the support for the interaction between Universal and Training suggests that learning was easier for monotone quantifiers than for non-monotone quantifiers, in line with our hypothesis. However, the interaction effect was more pronounced for the downward monotone quantifier vs. non-monotone quantifier pair than the upward monotone quantifier vs. non-monotone quantifier pair.

6.3.2 Convexity

For convexity, the mean intercept was β = -0.15 with the credible interval [CI: -0.37, 0.06] including zero, the Universal effect was $\beta = 0.51$ [CI: 0.22, 0.82], the Training effect was $\beta = 0.36$ [CI: 0.15, 0.56], and the interaction effect was β = -0.02 with the credible interval [CI: -0.32, 0.28] including zero. The Universal effect had $\Pr(\beta > 0) = 0.9995$, the Training effect $\Pr(\beta > 0) = 0.9995$, and the interaction effect $\Pr(\beta > 0) = 0.45$. The lack of interaction effect for convex vs. non-convex quantifiers suggests that they did not differ in learning rate. Nonetheless, participants were more accurate for the convex quantifier, as indicated by the Universal effect.

6.3.3 Quantity

For quantity, the mean intercept was $\beta = 0.68$ [CI: 0.48, 0.88], the Universal effect was $\beta = 0.13$ with the credible interval [CI: -0.15, 0.41] including zero, the Training effect was $\beta = 0.61$ [CI: 0.44, 0.78], and the interaction effect was β = 0.06 with the credible interval [CI: -0.18, 0.32] including zero. The Universal effect had $\Pr(\beta > 0) = 0.84$, the Training effect $\Pr(\beta > 0) = 1$, and the interaction effect $\Pr(\beta > 0) = 0.69$.

Because Steinert-Threlkeld and Szymanik (2019) found a greater effect for the pair of quantifiers *at least 3* and *the last 3* than *at least 3* and *the first 3*, we decided to test these two pairs in an additional analysis. For the pair *at least 3* vs. *the first 3*, we found that the mean intercept was $\beta = 0.91$ [CI: 0.61, 1.2], the Universal effect was $\beta = 0.12$ with the credible interval [CI: -0.27, 0.52] including zero, the Training effect was $\beta = 0.39$ [CI: 0.14, 0.63], and the interaction effect was $\beta = 0.07$ with the credible interval [CI: -0.28, 0.42] including zero. The Universal effect had $\Pr(\beta > 0) = 0.71$, the Training effect $\Pr(\beta > 0) = 0.9995$, and the interaction effect $\Pr(\beta > 0) = 0.66$.

For the pair *at least 3* vs. *the last 3*, we found that the mean intercept was $\beta = 0.46$ [CI: 0.23, 0.7], the Universal effect was $\beta = 0.54$ [CI: 0.18, 0.88], the Training effect was $\beta = 0.81$ [CI: 0.56, 1.05], and the interaction effect was β = -0.35 [CI: -0.71, -0.02]. The Universal effect had $\Pr(\beta > 0) = 0.99975$, the Training effect $\Pr(\beta > 0) = 1$, and the interaction effect $\Pr(\beta > 0) = 0.021$.

All three analyses showed that participants were more accurate for quantifiers that satisfied the quantity universal. However, the interaction effect was not in line with our hypothesis. Participants had comparable learning rates for *at least 3* than *the first 3*, and lower rates for *at least 3* than *the last 3*. Under our hypothesis we would expect the higher learning rate for *at least 3*.

6.3.4 Conservativity

Finally, for conservativity, the mean intercept was $\beta = 0.33$ with the credible interval [CI: -0.05, 0.7] including zero, the Universal effect was $\beta = 0.03$ with the credible interval [CI: -0.5, 0.56] including zero, the Training effect was $\beta = 1.17$ [CI: 0.91, 1.43], and the interaction effect was $\beta = 0.09$ with the credible interval [CI: -0.28, 0.44] including zero. The Universal effect had $\Pr(\beta > 0) = 0.54$, the Training effect $\Pr(\beta > 0) = 1$, the interaction effect $\Pr(\beta > 0) = 0.68$. The lack of pronounced interaction effect for conservativity supports our hypothesis. We did not expect participants to learn *not all* faster than *not only*.

In conclusion, the posterior probability of the Training effects was nearly 1. Therefore, we can confidently report an improvement in accuracy for all quantifiers. Moreover, we found a strong effect of interaction between Universal and Training effects for monotonicity. However the effect was stronger for downward entailing than upward entailing quantifiers. The Universal effect was strongest for monotonicity and convexity, and weaker for quantity (including zero in credible interval). The weakest effect was observed for conservativity.

6.4 Discussion

The goal of this study was to test the learnability hypothesis for three semantic universals in the domain of quantification. We predicted that participants would more quickly learn quantifiers satisfying the monotonicity (convexity), and quantity universals, but not conservativity.

Our predictions partially bore out. We found a robust effect of learnability for the monotonicity universal. We found that participants achieved higher accuracy at the end of the experiment for quantifiers (*at least 3*, *at most 2*) that satisfied the universal than quantifiers (*between 3 and 6*, *at most 2 or at least 7*) that did not. Moreover, participants were already more accurate for monotone quantifiers than non-monotone quantifiers in the first 25% of the trials and they learned monotone quantifiers faster.

In addition, our findings show more a complex relationship between learnability and convexity and quantity universals. We found that participants were already more accurate for the convex quantifier (*between 3 and 6*) than the non-convex quantifier (*at most 2 or at least 7*) at the beginning of the experiment, but we did not find the interaction effect. Moreover, participants were more accurate for quantifiers that satisfied the quantity universal (*at least 3*, *at most 2*) than quantifiers that did not (*the first 3*, *the last 3*) at the beginning of the experiment (however, the 95% credible interval included zero). The weakest effects were found for the conservativity universal. This finding is consistent with the hypothesis that the conservativity universal emerged under different pressures than learnability. We discuss these results in more detail below.

6.4.1 Monotonicity and Convexity

Our results contribute to ample evidence on the role of learnability in shaping monotonicity and convexity universals. Chemla, Buccola, and Dautriche (2019) found that adult participants learned monotone rules faster than non-convex rules. The difference between monotone rules and non-monotone but convex rules was not significant. Similarly, Steinert-Threlkeld and Szymanik (2019) showed that neural networks learned monotone than non-monotone quantifiers faster, yet, they did not test non-monotone but convex quantifiers.

Our study extends these findings by showing the difference between monotone and non-monotone but convex quantifiers, and non-monotone but convex and non-convex quantifiers. We demonstrated that quantifiers can be ordered according to learning difficulty, where monotone quantifiers are the easiest, non-monotone but convex quantifiers are more difficult, and non-convex quantifiers are the most difficult. This finding is consistent with the idea of *the degree of monotonicity* (Carcassi, Steinert-Threlkeld, & Szymanik, 2019). Carcassi, Steinert-Threlkeld, and Szymanik (2019) applied an information-theoretic measure of degrees that assigned value 1 to monotone quantifiers such as *some* and values close to 0 to highly non-monotone quantifiers, such as *an even number*. The quantifiers with higher degrees emerged via iterated learning. Although we did not apply the degree of monotonicity measure, we see a similar pattern in our results: quantifiers with higher degrees were easier to learn for participants.

We found the difference in accuracy between monotone and non-monotone and non-monotone and non-convex quantifiers already occurred in the first 25% of the trials. This means that when participants considered possible meanings of *gleeb* at the beginning of the experiment, they relatively quickly focused the hypothesis space on the monotone quantifiers. This effect can be explained by the higher frequency of monotone than non-monotone quantifiers (see Appendix D, Table D.2).

We also noticed that the upward monotone quantifier was already easier than the downward monotone quantifier at the beginning of the experiment. The polarity effect is well-established in verification experiments with upward vs. downward monotone quantifiers. Downward monotone quantifiers are more difficult to process (Just & Carpenter, 1971; Deschamps et al., 2015; Schlotterbeck et al., 2020; Szymanik & Zajenkowski, 2013) and it takes longer to verify them. This effect can be explained by so-called pragmatic accounts (Degen & Tanenhaus, 2019; Schlotterbeck et al., 2020). Participants disprefer downward monotone quantifiers in the production because they are less informative. This preference is also reflected in the frequency data (see Appendix D, Table D.2). Upward monotone quantifiers are more frequent than the downward monotone quantifiers. For example, the upward monotone quantifier construction in our experiment had higher *Zipf* value (5.19) than downward monotone quantifier (4.49). Pragmatic preferences could play a role in our experiment. Participants searched through

their hypothesis space starting with pragmatically preferred quantifiers.

The interaction effect between monotone and non-monotone quantifiers provides further support for this interpretation. Participants learned the monotone quantifiers faster. However, the interaction was stronger for the pair of downward monotone (*at most 2*) and non-monotone (*between 3 and 6*) quantifiers than upward monotone (*at least 3*) and non-monotone (*between 3 and 6*) quantifiers. The interaction effect showed that many participants already learned the upward monotone quantifier in the first 25% of trials and they did not improve their accuracy much during the experiment. For *at most 2*, participants needed more trials to learn the meaning correctly, because this quantifier is pragmatically dispreferred.

In contrast, we did not find the interaction for the convexity universal, meaning that the learning rate was the same for convex (*between 3 and 6*) and non-convex (*at most 2 or at least 7*) quantifiers. One reason for the lack of interaction could be that only a few participants reached the accuracy criterion of 83%. Many participants simply did not learn the quantifier, and this number was larger for non-convex quantifier. Other reasons could be that participants approximated the meanings of these quantifiers to a monotone quantifier. They improved in accuracy because they knew the correct answer in some of the trials, but they did not discover the meaning of the quantifier.

6.4.2 Quantity

Our data showed that the relationship between learnability and quantity universal is more complex. We found a moderate universal effect. Moreover, participants were more accurate for *at least 3* than *the first 3* and *the last 3*. Interestingly, we found that interaction effects went in different directions for *the first 3* and *the last 3*. The Training effect was comparable for the *at least 3* quantifier and the *the first 3* quantifier. Surprisingly, for *at least 3* and *the last 3*, we observed that the Training effect was greater for the non-quantitative quantifier. This finding suggests that for *at least 3* and *the first 3* participants came up with a close to correct hypothesis quickly and the Training effect was smaller at the end of the experiment. The accuracy for *the last 3* was lower than for *the first 3* in the first 25% of the trial. During the experiment, participants learned *the last 3* and achieved a similar accuracy to *the first 3* in the last 25% of the trials. Figure 6.3a illustrates the interaction patterns.

To summarize, this finding shows that *the last 3* was more difficult than *the first 3*, at least at the beginning of the experiment. The result shows a similar pattern to the neural networks finding (Steinert-Threlkeld & Szymanik, 2019). Neural networks learned *the last 3* slower than *the first 3*. This finding cannot be explained by universal properties. Moreover, it is also rather implausible that *the last 3* is more difficult due to the difference in frequency, because both quantifiers have very similar *Zipf* values (3.70 and 3.86, see Appendix D, Table

D.2). We propose that the difference in learning performance could be due to salient properties of stimuli. The triangles at the beginning of the row could have attracted participants' attention more than the triangles at the end. When testing the quantity universal, other pressures on cognitive systems, such as visual salience, should be considered in future studies.

6.4.3 Conservativity

We did not find support for the role of learnability in shaping the conservativity universal. We found that the accuracy for *not all* was the same as for *not only* at the beginning of the experiment. This means that the hypothesis that *gleeb* is a conservative quantifier was not more likely than the hypothesis that it is a non-conservative quantifier. In contrast to previous studies (Hunter & Lidz, 2013; Spenader & de Villiers, 2019; Steinert-Threlkeld & Szymanik, 2019), we used Bayesian statistics instead of a frequentist approach. In this way, we were able to tested the hypothesis about lack of effect. We showed that the Universal effect for conservativity was smaller than for the other universals. Moreover, we did not find strong evidence for the interaction effect. Consistent with most of the previous studies (Spenader & de Villiers, 2019; Steinert-Threlkeld & Szymanik, 2019), we conclude that conservativity should be explained by other pressures than learnability.

The structural account for conservativity (Romoli, 2015) provides an alternative explanation for the conservative universal. According to this approach, non-conservative quantifiers create trivial meanings or meanings equivalent to conservative quantifiers. Conservativity as the semantic universal emerged under the pressure of the syntax-semantics interface (Romoli, 2015). Our experiment did not test this account; therefore, we cannot conclude that our data support the syntax-semantics interface hypothesis. We leave the evaluation of this explanation to future studies.

6.4.4 Methodological remarks

Having discussed the main findings, we want to take stock of a few methodological issues related to this study. In comparison to previous experiments, we introduced several methodological advancements. We adopted the same experimental paradigm to different universals. We tested the differences between quantifiers that do and do not satisfy the universals as well as differences between universals. We assumed that, if we found a difference between quantifiers within one universal, but not within another, then the difference between universals would likely to be due to different properties of universals rather than the different experimental paradigm. Additionally, we included the quantity universal, which was not previously tested in humans.

Moreover, in contrast to rather simple previous designs,[12] we designed an experiment with a high variation of stimuli. The variation in the number of objects excluded the possibility of participants memorizing the correct responses rather than learning the meaning of the quantifier. It also created a more naturalistic setting.

We noticed that our experiment was challenging. About one-third of the participants were unable to learn quantifiers and performance systematically above chance. For comparison, Chemla, Buccola, and Dautriche (2019) found that participants learned the most difficult, non-convex rules, in an average of 91 trials. In contrast, our experiment consisted of 96 trials, and only 2 out of 30 participants learned the non-convex quantifier. We decided against adding even more trials because we did not see more improvement in accuracy in our pilot studies after extending the length of the experiment.

We suspect that the experimental design used by Chemla, Buccola, and Dautriche (2019), with a fixed number of objects, helped participants constrain the space of all possible quantifiers under consideration. Hence, they learned faster than in our experiment. Altogether, our study shows that changes in the design (e.g., variation in the number of stimuli) can increase difficulties and reveal subtle differences between quantifiers.

Besides changes in experimental design, we also addressed the replication problem of previous studies. Two studies that investigated the learnability effect for conservativity in humans (Hunter & Lidz, 2013; Spenader & de Villiers, 2019) were highly underpowered. Hunter and Lidz (2013) recruited 10 participants per quantifier, whereas in a replication experiment, Spenader and de Villiers (2019) included only 9 participants per quantifier. Both studies also had a very short training block, hence participants might not have had enough time to become familiarized with the experiment and learn the quantifier at the same time. This means that the test data may have contained a substantial noise unrelated to the tested universal (for example, because participants could have made many response errors). As a consequence, these studies showed inconsistent results.

Studies with a very small sample size might lead to inconsistent results (cf. Aarts et al., 2015). One advantage of online experiments is the possibility to collect a large amount of data in a short time (Kochari, 2019). We used this opportunity to collect substantially larger sample sizes per quantifier and obtain more reliable results. We also included longer training to reduce the noisiness of the data.

[12] We tested the experimental design in several pilot experiments. We also tried a design with a fixed number of figures ($N = 8$). To avoid a ceiling effect and very fast learning of monotone quantifiers, we decided to introduce variation in the number of objects.

6.4.5 Potential confounds

Needless to say, adult participants are already affected by the experience with their native language. This creates potential confounds in our study. Although eliminating all of these confounds is not possible, we put effort into minimizing their effect.

Quantifiers' frequency

The first confound comes from the quantifier frequency. Quantifiers that satisfy universals are easier to learn, and as a result, they are also used more frequently in languages. The frequency of usage can in turn affect the speed of learning, because participants will come up with the more frequent quantifiers as possible hypotheses faster than with the less frequent quantifiers. If such a relationship in the language exists, it cannot be simply overcome. To control for this confound, we checked the frequency of quantifiers used in our experiment (see Appendix D, Table D.2).

We checked the frequencies of quantifiers in our experiment in the Corpus of Contemporary American English (Davies, 2008-). We used *Zipf* values (van Heuven et al., 2014) to compare the frequencies. We used the *Zipf* scale because it is a standardized scale and gives an intuitive interpretation of values. The *Zipf* values of 1 to 3 correspond to low-frequency words, and *Zipf* values between 4 and 7 to high-frequency words. Most of the quantifiers in our study had frequencies around 3 to 4 *Zipf* values, meaning that they had a moderate frequency. One quantifier, *at most 2 or at least 7*, did not occur in the corpus. The quantifier *between 3 and 6* has *Zipf* value around 2, which classifies this quantifier as a low-frequency quantifier. It is worth mentioning that the construction containing between NUM and NUM had a moderate frequency. In contrast, *not only* had a *Zipf* value above 5, meaning that it is a highly frequent expression. Overall, we believe that our choice of quantifiers, except *at most 2 or at least 7*, was balanced in terms of frequency as much as possible.

The relationship between universality and frequency resembles a chicken and egg problem. On the one hand, the quantifiers that satisfy the universals should be easier to learn and therefore more frequently used and finally even lexicalized. Our analysis showed that lexicalized quantifiers such as *all*, *some*, *many*, *few*, and *most* are the most frequent (see Appendix D). On the other hand, more frequent expressions are also faster to acquire. The formal theory of quantifiers does not predict the causal relationship between universality and frequency, and our experiment cannot fully address this confound. One neural network experiment (Mhasawade, Szabó, Tosik, & Wang, 2018) gives partial insight into the effect of frequency on quantifier learning. Mhasawade et al. (2018) demonstrated that even when the training data distribution favored conservative quantifiers, there was no learnability bias toward conservativity. Although there is a gap in the

extrapolation of neural network behavior to humans, this finding shows that universal properties are not simply the effect of the frequency of quantifiers.

Lexicalization abilities

The second objection to our study could be that our participants already spoke a language with lexicalized quantifiers that satisfied universal properties. Therefore, what we tested on our participants is not the learnability *per se* but their ability to replace a new word with some other words known in their native language. This task would naturally be more difficult for quantifiers that do not satisfy universals, because they are not lexicalized. To control for this confound, we chose only non-lexicalized quantifiers. Therefore, what we tested in our experiment was not just the ability to replace the new word with a corresponding single word in participants' native language, but rather the ability to lexicalize a complex quantifier as a single word.

Word vs. rule-learning paradigms

In our experiment, participants learned an artificial determiner instead of a rule. We made this design choice, because we intended to capture more specific linguistic bias rather than more general learning of concepts (cf. Chemla, Buccola, & Dautriche, 2019). We used a word-learning experimental design in which quantifier meanings were lexicalized as a new artificial word *gleeb* (Hunter & Lidz, 2013). In contrast, Chemla, Buccola, and Dautriche (2019) instructed participants to learn a rule. They claimed that participants did not define the rule in terms of quantifiers. Moreover, some participants had a problem with the explicit formulation of the rule. According to the researchers, the experiment captured a more general cognitive bias toward convexity.

Nonetheless, the objection could arise that our participants did not recognize the new words as a determiner. If this were the case, then our experiment would be similar to a rule-learning experiment. To control for this confound, we asked participants to indicate what they thought *gleeb* meant. While some participants proposed a quantifier (e.g. like "first three" for *the first 3*, "some but not all" for *not all*, "minority" for *at most 2*), others proposed a rule (e.g., "something to do with blue triangles being present" for *not all*, "three in a row on the right side" for *the last 3* or "three or more red triangles." for *at least 3*, "3, 4, 5 or 6 red triangles" for *between 3 and 6*). There were also many "I don't know answers" or answers that were difficult to interpret. While wee did not systematically classify typed responses into any categories, we think that our participants may have applied the mixed rule and word-learning strategies.

Further studies should focus on methodological differences between word vs. rule-learning paradigms. We considered two possible solutions to reinforce word-learning instead of rule-leaning. However, we did not find them satisfactory. The

first solution would be to train participants in our task and test them in a task with a different set of objects to see whether they could extrapolate the learned quantifier. However, it is also possible to extrapolate the learned rule, so this manipulation would not completely overcome the problem. The second solution would be to vary the sentence in each trial. In our experiment, the sentence always referred to red triangles. We could, however, use various types of objects and vary the sentence's scope. This manipulation would bring more attention to the sentence, but would not guarantee that participants did not represent the task as a rule and not as a determiner.

The specific linguistics biases are likely shaped by more general cognitive biases. The origin of cognitive bias toward convexity seems to have a more primary root than in human language preference. Chemla, Dautriche, Buccola, and Fagot (2019) showed that baboons (*Papio papio*) learn the convex rules more easily than non-convex rules. Moreover, in the domain of content words, Gärdenfors (2000) formulated a convexity universal for color: all color terms denotate convex regions of color space. Steinert-Threlkeld and Szymanik (2020) showed that the degree of convexity predicts the neural network's ability to learn color term categories.

Convexity also plays an important role in learning concepts. The conceptual space is organized in such a way that objects that share properties are more likely to be labeled with one category. Metaphorically speaking, the concepts should not have gaps. For example, a set containing husky, labrador, and chihuahua forms a convex category of dogs, and a set containing dog, cat, and hamster forms a convex category of pets. Adult participants, when presented with a new word in the artificial language and example category members (like husky, labrador, and chihuahua), will generalize this new word to other members of the category (like beagle) at the same category level (Xu & Tenenbaum, 2007). However, when participants are presented with a set of objects that do not form a convex category, (e.g., a set containing dog and tree), they will not generalize the new category label to all objects falling under the broader category (e.g., living things), but rather treat the new word as homophony (Dautriche & Chemla, 2016). Similarly, linguistic behavior and preference for convex categories was observed in children (Dautriche, Chemla, & Christophe, 2016).

While convexity seems to be a more general cognitive bias, conservativity and quantity are more language specific. This could explain why the effects for convexity and monotonicity were more pronounced in our data. Even if participants learned the rule instead of the quantifier, they were still affected by the bias. Nonetheless, we believe that our findings are not just an experimental design artifact. Further studies should replicate this effect in different experimental set-ups to test the linguistic specific vs. general cognition nature of the semantic universal.

6.4.6 Minimal pair methodology limitations

Finally, the experimental testing of semantic universals poses practical difficulties. While simulation studies can have many quantifiers with different degrees of semantic universals (Carcassi, Steinert-Threlkeld, & Szymanik, 2019), the experimental studies are doomed to use the minimal pairs methodology. The minimal pair methodology assumes that the investigated quantifiers are representative of the semantic universals and that they are comparable for other properties. In our experiment, we tested quantifiers that were extensively tested previously (Hunter & Lidz, 2013; Chemla, Buccola, & Dautriche, 2019; Spenader & de Villiers, 2019; Steinert-Threlkeld & Szymanik, 2019).

Being aware of the methodological limitations, we stress that to draw robust conclusions about the role of learnability in shaping semantic universals, we need evidence from many sources. Previous studies provided evidence in neural networks (Mhasawade et al., 2018; Steinert-Threlkeld & Szymanik, 2019), children (Hunter & Lidz, 2013; Spenader & de Villiers, 2019), animals (Chemla, Dautriche, et al., 2019), language evolution (Carcassi, Steinert-Threlkeld, & Szymanik, 2019), and simulation experiments (van de Pol et al., 2019). Our study adds a building block to this picture. It consistently shows that learnability plays a role in forming semantic universals, albeit to a different extent for different universal properties. We cannot, however, conclude that learnability pressure is the only or most important form of pressure. Other pressures, such as complexity (van de Pol et al., 2019, 2021), informativeness (Carcassi, Schouwstra, & Kirby, 2019), the abilities of the visual system, and syntax-semantics interface (Romoli, 2015), can also contribute to the formation of semantic universals. Moreover, the universals might arise from trade-offs between different pressures. For example, the complexity-simplicity trade-off shapes semantic categories of content words such as kinship terms (Kemp & Regier, 2012) and function words, e.g., quantifiers (Steinert-Threlkeld, 2020).

6.4.7 Conclusions

The goal of the current study was to evaluate the learnability hypothesis in the realm of quantifiers. This hypothesis states that quantifiers satisfying semantic universals are easier to learn. We tested three well-established semantic universals (Barwise & Cooper, 1981; Peters & Westerståhl, 2008): monotonicity (convexity), quantity, and conservativity. We provided evidence for a noticeable role of the learnability effect in monotonicity. Learnability was found to play an intricate role in the quantity universal. We did not observe the learnability effect for conservativity. Together with other evidence (van de Pol et al., 2019; Spenader & de Villiers, 2019; Steinert-Threlkeld & Szymanik, 2019), this leads us to the conclusion that conservativity emerged under a different form of pressure than other semantic universals.

Chapter 7

Conclusions

The goal of this thesis was to propose a cognitive model of quantifier representation and verification. To this end, I conducted behavioral and electroencephalography experiments and applied three different computational models to the data from the quantifier verification task. In the last chapter of this thesis, I will take stock of the main findings and present the cognitive model. I will discuss the status of quantifier thresholds and the formal approaches to model uncertainty about the thresholds. Moreover, I will address inconsistency in modeling results of the polarity effect. Finally, I will discuss the advantages and disadvantages of each of the computational models.

7.1 Summary of the main findings

One of the challenges for the logical models of quantifiers was to account for individual differences in meaning representations. In Chapters 2 and 3 of this thesis, I investigated the ranges of individual differences in quantifier representations and proposed two computational models to account for the between-participants variability in thresholds. My experiments and modeling showed that quantifiers with sharp meaning boundaries (e.g., *more than half* and *fewer than half*) have similar representations for all participants, while vague quantifiers (e.g., *many* and *few*) are sensitive to individual differences. Moreover, I established an intermediate class of quantifiers (e.g., *most*). According to the Generalized Quantifier Theory, *most* has a fixed meaning, but the experimental findings showed that its thresholds vary between participants. I also demonstrated that individual differences are stable over time.

The individual differences in quantifier representations go beyond differences in the truth conditions. In Chapter 2, I showed that participants create clusters with different meanings and orders of quantifiers on a mental line. Moreover, some quantifiers (*many*) had a more flexible position on the mental scale.

In addition to modeling the variability in thresholds, I proposed two computational models that quantified the level of uncertainty about quantifier meanings. Using computational modeling, I was able to disentangle the vagueness of the quantifier and the variability in truth conditions between participants. At first glance, it seems that vagueness and variability in truth conditions should be correlated. Nonetheless, they are conceptually different semantic properties. The variability in truth conditions reflects how the groups of participants disagree about the meaning of quantifiers. Vagueness, in turn, indicates a within-participant level of certainty about the representation.

The experiments in Chapters 2 and 3 contributed to understanding the ranges of variability in the semantic representations of logical words. Two independently motivated computational models showed similar findings. The consistency of the result across different analyses speaks for its robustness. The next step to fully understand the individual differences in quantifiers would be to investigate the origin of the individual differences. A number of different hypotheses can be formulated here. Firstly, individual differences could be a result of sociolinguistic factors, e.g., education level (Verheyen & Storms, 2018) or participants' traits (Verheyen et al., 2018). Secondly, they may lay in the domain-general or linguistic-specific cognitive abilities of participants. Certainly, the processing of quantifiers can be linked to working memory and executive functions (Just & Carpenter, 1992; Kidd et al., 2018; Zajenkowski & Szymanik, 2013; Zajenkowski et al., 2014), and pragmatic abilities (Nieuwland et al., 2010). The third possibility would be to link the individual differences to personality traits, for example, sensitivity to sensory input (Heim et al., 2020). Finally, it is also possible that the individual differences are a result of the combined effect of these three factors. This could be tested in future work.

In this thesis, I studied quantifier representations in the verification task. In this task, participants assessed the truth value of the quantified sentence based on numerical information provided in the form of a sentence or picture. I modeled the evaluation of the quantifiers as a process of comparison between the internal representation (threshold) and numerical information. The comparison process was affected by the properties of quantifiers (e.g., vagueness, polarity, and threshold) and task-specific information (e.g., percentages, numerosity of sets, and the proportion of objects). In Chapter 3, I modeled the speed of quantifier verification as dependent on semantic features of quantifiers and showed that the verification of the vague quantifiers is proportion-dependent.

In Chapters 3, 4, and 5, I demonstrated that the polarity of quantifiers (negative vs. positive) affects the speed of verification. In Chapters 4 and 5, I tested two different explanations of this effect: the two-step model and the pragmatic account. In light of these findings, the pragmatic account is a more likely explanation. Negative quantifiers are more difficult to evaluate because they hinder the process of building expectations about the sentence.

Finally, in Chapter 6, I tested the link between participants' cognitive abilities

and the semantic universals, namely the learnability hypothesis. I demonstrated that participants uncovered the meaning of a new quantifier faster depending on its formal properties. For example, monotonicity eased the learning of quantifiers, while conservativity did not. Steinert-Threlkeld and Szymanik (2019) used a heat map metaphor to describe this phenomenon. While participants discover a new meaning, they have to choose from a wide range of hypotheses. Quantifiers are more or less likely to be selected depending on the formal properties because they are "hotter" on the heat map. According to the heat map metaphor, "hot" quantifiers are easily accessible to participants, while "cold" quantifiers are difficult to access. Moreover, the heat map also has different shades of hot. For example, the monotone quantifiers are hotter than non-monotone quantifiers, and the upward monotone quantifiers are the hottest. The learnability effect is a result of the heat distribution on the map. In conclusion, this finding showed that the learnability hypothesis can explain some of the semantic universals.

7.2 Cognitive model

In the introduction to this thesis, I contrasted the logical view on quantifiers with a cognitive view. I demonstrated that the logical view is insufficient to account for psycholinguistic phenomena related to the verification of quantifiers. Specifically, the logical view does not allow for individual differences between participants. Moreover, it does not propose a link between the formal properties of quantifiers, verification, and constraints on humans cognitive abilities. The cognitive model informed by the experimental data presented in Chapters 2 to 6 fills these gaps.

The cognitive model proposes that participants have a representation of each quantifier stored in memory in the form of an internal threshold. Thresholds depend on the truth-conditional representation of the quantifier, and also on the linguistic experience of participants. Moreover, participants have different levels of certainty about thresholds. The source of uncertainty might lay in the semantics of a quantifier (e.g., vagueness), but also may be specific to the individual. While performing the verification task, participants compare their internal representations to the experimental input. The speed of this comparison process depends on various factors, such as the input properties, threshold, the polarity of the quantifier, or vagueness.

The thresholds are ordered on the mental line. Moreover, quantifiers differ in the accessibility of their representation to participants. The accessibility of their representation depends on the formal properties of quantifiers, such as polarity or whether they satisfy universals. The more accessible quantifiers are verified faster and easier to acquire.

The cognitive model of quantifiers postulates individual differences in meaning representations. So far, the model has not questioned whether quantifiers have truth-conditional representations. In light of the findings presented in this thesis,

it seems that the truth-conditional representations derived from the Generalized Quantifier Theory are not sufficient to account for variability in thresholds and vagueness (cf. Glöckner, 2006). To account for the data presented here, we may need to take a step beyond the Generalized Quantifier Theory. In a similar vein, van Tiel et al. (2021) showed that the Generalized Quantifier Theory has to be enriched with a pragmatic module to account for linguistic behavior in quantifier production experiments. The further modification of the Generalized Quantifier Theory is one possible direction to take. Another approach would be to find an alternative, formal framework to model the flexibility and uncertainty about meanings. For example, Chater and Oaksford (1999) proposed the probabilistic semantics of quantifiers. The theory of fuzzy quantifiers, in turn, based the quantifiers' semantics on the fuzzy set theory (Glöckner, 2006). Both approaches might be considered as alternatives to the Generalized Quantifier Theory.

7.2.1 Probabilistic semantics of quantifiers

The starting point for the probabilistic semantics of quantifiers is the observation that human cognitive abilities (e.g., reasoning) evolved as an adaptation to the uncertainty of the real world. The logical models cannot properly explain the use of language (e.g., everyday reasoning) because they do not account for uncertainty. To fill this gap, Chater and Oaksford (1999) proposed applying a probabilistic approach based on probability theory instead of logic.[1]

Chater and Oaksford (1999) developed a probabilistic approach to syllogistic reasoning, called the Probability Heuristics Model. The model included a set of reasoning heuristics and probabilistic semantics of quantifiers (e.g., *most* and *few*). According to the probabilistic approach, quantifiers are ordered in terms of their informativeness, which is defined as surprisal. To further analyze the informativeness of quantifiers, Chater and Oaksford (1999) proposed treating the quantified sentences as probabilistic statements. quantifier's semantics are defined in terms of conditional probability. For example, the statement "*All* Xs are Y" is defined as $P(Y|X) = 1$, meaning that the probability of Y given X is equals 1. *Most*, in turn, is defined as $P(Y|X) = 1-\Delta$ (where Δ is small), meaning that the probability of Y given X is less than 1, but still high. The model also takes into account the polarity of quantifiers, which Oaksford, Roberts, and Chater (2002) categorized as a pragmatic property.

To summarize, the Probability Heuristics Model begins with the assumption that logical models can not explain human reasoning because they can not effectively model uncertainty. To account for the patterns of experimental data, Chater and Oaksford (1999) proposed a model based on probability theory instead of logic. While they made a correct observation that classical, two-valued logic can not model uncertainty well, their shift to probability theory is not a

[1] The authors refer to logic as two-valued classical logic.

necessity. The two-valued logic of the Generalized Quantifier Theory is not the only possible logical framework for quantifiers. The theory of fuzzy quantifiers provides an alternative approach to model vagueness and uncertainty without referring to probability theory.

7.2.2 Fuzzy quantifiers

In contrast to probabilistic semantics, the fuzzy theory of quantifiers proposed modeling uncertainty by extending the logical theory rather than reaching to probability theory. The idea of fuzzy quantifiers is based on the fuzzy set theory developed by Zadeh (1983). The motivation for the development of this theory was the observation that vagueness is an inherent property of natural language (Glöckner, 2006). Moreover, vagueness is a desired property for communication purposes. Vague language is more tolerant of the imperfection of human cognitive abilities and more flexible. Vagueness should be distinguished from context dependency. Context-dependent expressions allow for multiple comparison classes and, therefore, can have many alternative interpretations. However, once the context is fixed, the alternatives are constrained to one comparison class. In contrast, vagueness does not disappear in a fixed context. In other words, even in a fixed context, vague expressions have borderline cases.

The theory of fuzzy quantifiers rejects the two-valued logic that constrains the Generalized Quantifier Theory. The Generalized Quantifier Theory does not allow for the vagueness of quantifiers' arguments or a gradual quantification, and therefore can not model vague expressions (Glöckner, 2006). To account for vagueness, it has been proposed to adopt the continuous-valued model instead of the two-valued model (see Glöckner, 2006). The continuous-valued model can be constructed on the basis of the fuzzy set framework. In this framework, the quantifier is not a mapping from sets into true or false values ($\{0, 1\}$), but into the interval ([0, 1]).

The fuzzy theory of quantifiers assumes two sources of fuzziness in quantified statements. The quantifiers themselves constitute the first source. Some quantifiers have precise (or *crisp*) meanings (e.g., *every*), while others have *approximate* meanings (e.g., *many*). Moreover, both types of quantifiers can take fuzzy or crisp input arguments (restrictor or scope). The category of fully fuzzy quantifiers consists of the approximate quantifiers with fuzzy arguments (also known as Type IV quantifiers, see Y. Liu & Kerre, 1998). The semi-fuzzy quantifiers are fuzzy in terms of gradual quantification but have crisp inputs. The semi-fuzzy quantifier theory can be treated as an extension of the Generalized Quantifier Theory because all generalized quantifiers can be defined as semi-fuzzy quantifiers.

Based on the fuzzy logic approach, van Tiel et al. (2021) proposed semantics of quantity words based on the prototype theory. According to this account, each quantifier has a prototype meaning, which reaches the maximum true value. The truth value of the quantifier in a given situation can vary as a function of a distance

from the prototype meaning. The prototype-based has been tested against the model based on Generalized Quantifier Theory truth-conditional semantics in a production experiment (van Tiel et al., 2021). The truth-conditional model could explain the data equally well as the prototype-based model, but only if it was enriched with a pragmatic module.

7.2.3 Modeling of thresholds and vagueness

The probabilistic semantics of quantifiers and fuzzy quantifier theories provide more flexible accounts than the Generalized Quantifier Theory. Can these accounts provide a better theoretical framework to the cognitive model of quantifiers than the Generalized Quantifier Theory?

In the light of the findings by van Tiel et al. (2021), the truth-conditional semantics model based on the Generalized Quantifier Theory can be defended if it is enriched with the pragmatic module. However, there are two crucial differences between the approach used by van Tiel et al. (2021) and the cognitive model presented in this thesis. Firstly, van Tiel et al.'s (2021) model is a production model, while the cognitive model presented here is a verification model. During the production of the utterance, participants take into account the meaning of the quantifier, but also its pragmatic value e.g., informativeness. The assumption about the pragmatic abilities of participants was crucial to improving the fit of the truth-conditional semantics model in the van Tiel et al. (2021) study. As a consequence, van Tiel et al. (2021) had to make an assumption about the relationship between quantifiers (e.g., which quantifier is more informative). In contrast, the cognitive model presented in this thesis does not assume any relationship between quantifiers.

Secondly, van Tiel et al. (2021) argued that the vagueness of the quantity words, such as *few* and *many*, can be derived from a non-linguistic source, for example, imprecise representations in the Approximate Number System (Dehaene, 1997). In contrast, in this thesis, I showed that vagueness and variability in truth conditions of quantifiers can be observed even in a task in which participants are unlikely to use the Approximate Number System. Therefore, I argue that vagueness is a property of a quantifier and not a by-product of the verification process.

The probabilistic semantics approach also captures vagueness as a part of semantic representations by referring to the probability theory. However, Glöckner (2006) provided convincing arguments against the probabilistic treatment of vagueness. The expectations are the core concept of the probability theory. In other words, probability theory refers to the epistemic states. It can be used to model context dependency, for example, by assuming that the different comparison classes give rise to different expectations. Moreover, the expectations can be updated once new information is available. Vagueness, in turn, cannot be resolved even with very precise information. One of the characteristics of vagueness is

the borderline cases, which remain controversial with respect to truth value even when more information is provided. Moreover, probability theory often refers to surprise as a measure of expectations. The borderline cases are by definition undefined in terms of truth value and the expectations about them can be neither revised nor confirmed. Vagueness cannot be surprising.

In the computational models presented in this thesis (the three-parameter logistic model and the Diffusion Decision Model), vagueness was implemented as part of the semantic representation of a quantifier, namely as a scale parameter of the logistic function. In this way, the additional assumption about the Approximate Number System representations was not needed. Moreover, both thresholds and vagueness were defined without referring to probability theory. The semantic representations proposed in the thesis are different from than the probabilistic semantics and the enriched Generalized Quantifier Theory truth-conditional semantics proposed by van Tiel et al. (2021).

The semantic representations investigated in this thesis are also different from the prototype-based semantics (van Tiel et al., 2021). Moreover, the experimental data presented in Chapter 3 speaks against the prototype-based model. Based on the prototype-based model, one can formulate a prediction about the reaction times in a verification task, namely that the verification should be faster when the verified proportion is close to the prototypical case. Specifically, the prototype model predicts a non-linear relationship between the reaction times and proportion. Reaction times should be slower around the threshold, then faster closer to the prototype, and then slower when the proportion is again further from the prototype. In contrast, I found that the reaction times for *most*, *many*, and *few* were proportion-dependent, but they were slower around the threshold and faster as the proportion was further from the threshold. Crucially, this relationship was linear. Consistently with the conclusions of van Tiel et al. (2021), I reject the prototype-based model.

In conclusion, in this thesis, the semantic representations of quantifiers are truth-conditional, but they go beyond the Generalized Quantifier Theory representations. By introducing the vagueness parameter, I allowed for a fuzziness of the representations around the threshold. The additional advantage of the model presented in the thesis is that it captures participants' behavior with respect to the polarity of the quantifier.

7.3 Polarity effect

In Chapter 4, I tested two competing accounts explaining the polarity effect: pragmatic and two-step models. The Diffusion Decision Model results supported predictions of both accounts. In Chapter 5, I tested the two-step model more precisely by estimating the number of processing stages for negative (*fewer than half*) and positive (*more than half*) quantifiers. The hidden semi-Markov model

multivariate pattern analysis did not support the two-step model. The conclusions of theses two chapters were inconsistent.

The interpretation of the non-decision time differences in Chapter 4 should be revised based on the findings from Chapter 5. The difference in the non-decision times parameter was not due to by the extra processing step as initially assumed. The question that remains, however, is what kind of cognitive processes are responsible for the difference. Both studies have shown two sources of differences in reaction time data between positive and negative quantifiers.

One possible explanation is that the difference in non-decision time from Chapter 4 was reflected by the difference in the last stage of processing in Chapter 5. Another possibility is that the difference in non-decision time was due to the combined duration of several stages (cf. Berberyan et al., 2020). To answer this question, further studies are needed to test the link between the Diffusion Decision Model parameters and stages of processing. The methodological lesson learned from these studies is that we should investigate linguistic phenomena using various computational models to draw firm conclusions.

In addition, the polarity effect was also present in the learnability study in Chapter 6. This finding was rather unexpected because the monotonicity universal does not predict the difference between upward and downward entailing quantifiers. However, the learnability effect might also be due to factors other than universals properties, for example, lower frequency of negative quantifiers or pragmatic preferences of participants.

For future directions, I suggest studying the polarity effect using the Diffusion Decision Model and hidden semi-Markov model multivariate pattern analysis together to better understand the relationship between the models' parameters. Moreover, it would be fruitful to test other pairs of polar opposite expressions, such as adjectives. In this thesis, I used the notions of polarity and monotonicity interchangeably because the tested quantifiers differed in both properties at the same time. However, by contrasting quantifiers with adjectives, Agmon et al. (2019) showed that polarity and monotonicity are two different sources of processing difficulties. Based only on data on quantifiers, it is not possible to make this distinction. Further studies could include adjectives to disentangle polarity and monotonicity and test their independent effects on processing stages.

7.4 Relationship between computational models

In this thesis, I applied three computational models to behavioral and electroencephalography data. The models differed in terms of theoretical commitments to the interpretation of parameters (model based on some cognitive theories vs. data-driven model), the level of detail in measuring cognitive processes (model of only choices vs. model of processing stages), and the number of parameters

to estimate. The levels of model description are not aligned, and the three computational models implemented in this thesis are orthogonal concerning these dimensions.

The three-parameter logistic model was the most parsimonious computational model. The interpretation of the model parameters was loosely inspired by the Item Response Theory. However, the model does not enforce commitments to cognitive theories. The main disadvantage of the logistic model is that it does not account for the processing of quantifiers, as it is a model of choices only. Nonetheless, the model was sufficient to capture the differences in thresholds and vagueness between quantifiers and to account for individual differences. It also captured the basic measure of participants' behavior, namely response errors.

The Diffusion Decision Model is a canonical model of processing. It accounts for both choices and reaction times, which is its main advantage over the three-parameter logistic model. The Diffusion Decision Model has clear cognitive interpretations of parameters. On the one hand, this could be seen as an advantage over the logistic model because the linking between parameters and cognitive processes is less arbitrary. On the other hand, the model is not completely theory-free and requires some commitments to theoretical assumptions, for example, to the evidence accumulation framework.

The main constraint of the Diffusion Decision Model is that it interprets the decision-making process at a fairly abstract level. It assumes only two stages of processing: decision and non-decision stages. According to the model assumptions, the non-decision stage is a sum of several stages such as perceptual processing or response. The model, however, does not distinguish these processes by employing different parameters.

The Diffusion Decision Model provides a somewhat simplistic view of processing. Berberyan et al. (2020) have shown that both decision and non-decision stages can be split into several sub-stages. Moreover, recent studies (e.g., Scaltritti, Job, Alario, & Sulpizio, 2020) questioned the sharp difference between decision and non-decision (motor response) stages in the processing of linguistic input. The level of analysis of the Diffusion Decision Model does not allow zooming in on processing stages in the verification task. For example, to test the two-step processing hypothesis, the more fine-grained measure of stages was needed.

The hidden semi-Markov model multivariate pattern analysis goes a step further than the Diffusion Decision Model, and provides a detailed model of processing stages. The hidden semi-Markov model multivariate pattern analysis is the most theory-free and data-driven model. It links the processing stages to theories of electroencephalography signal generation (Anderson et al., 2016), but it does not carry any commitments to specific cognitive theory. As a consequence, it allows for high degree of freedom in the interpretation of the modeling results, but also poses a risk of falling into speculation in interpretations.

Finally, I used different methods to estimate the models' parameters. The

parameters of the logistic model were estimated using the hierarchically Bayesian approach, the parameters of the Diffusion Decision Model, and the hidden semi-Markov model multivariate pattern analysis via the maximum likelihood estimation. The difference in fitting methods is not a specific characteristic of the models. For example, the parameters of the Diffusion Decision Model can also be estimated via the Bayesian approach (see Boehm et al., 2018, for discussion of different approaches). Importantly, the choice of fitting strategy did not affect the results obtained. For example, both the hierarchically Bayesian three-parameter logistic model and the maximum likelihood Diffusion Decision Model gave robust results on variability in threshold and vagueness.

To conclude, each of the computational models applied provided a unique insight into the representation and verification of natural language quantifiers. The choice of the specific model should always be motivated by the theoretical commitments involved in the model and the nature of the modeled phenomenon. Further studies should apply various computational models to study the single phenomenon in order to provide robust results and to compensate for the shortcomings of each model.

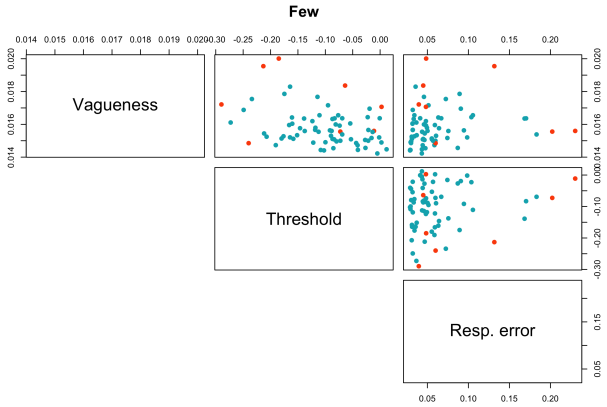
7.5 Coda

In this thesis, I presented the cognitive model of quantifier representations. The model was supported by behavioral and neural data from several quantifier verification experiments and modeling results from three computational models. My findings challenge the dominant formal approach to quantifiers and call for a new framework that accounts for individual differences in representation and verification of quantified sentences.

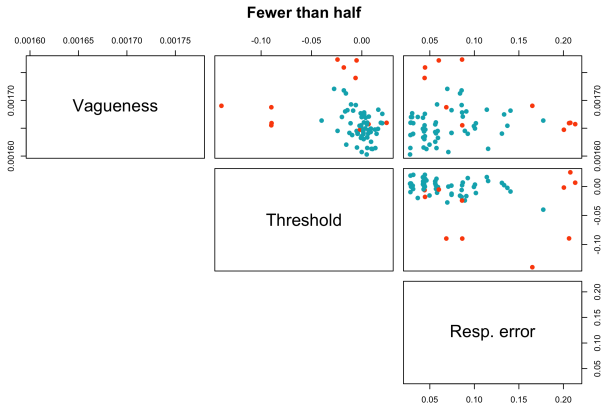
Appendix A

Appendix to Chapter 2

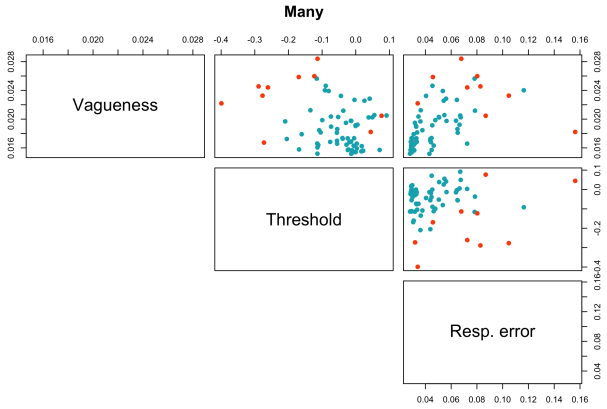
Figure 11 illustrates how relationships between model parameters for each quantifier are affected by influential observations. We computed the Cook's distance using the *ols plot cooksd bar* R function in the package *olsrr* (Hebbali, 2020).



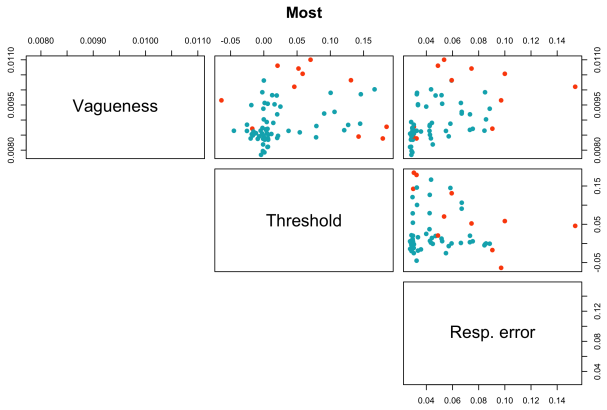
(a)



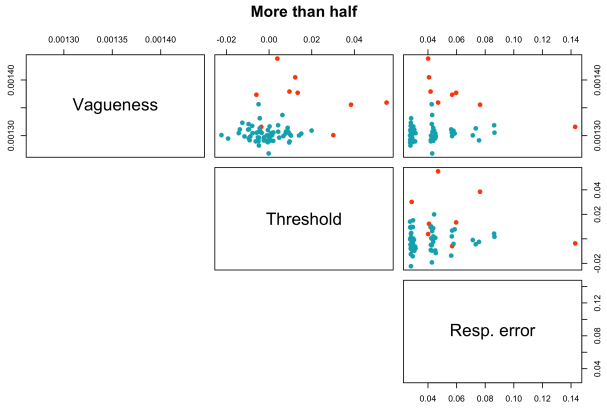
(b)



(c)



(d)



(e)

Figure A.1: The scatter plots illustrate the relationships between model parameters (abbreviation Resp. error - response error) for each quantifier. The influential observations according to Cook's distance are indicated in red.

Appendix B

Appendix to Chapter 3

B.1 Additional analyses

B.1.1 DV hypothesis testing - regression model comparison

To test the effect of proportion on reaction times for *most* in comparison to *more than half*, we tested the random structure of the mixed-effects regression model. In the first step, we included by-subject and by-item random intercepts. While participants were likely to vary in reaction times, we expected the small random effect of item, because we used pseudowords. We found that the model with by-subject and by-item random intercepts did not significantly improve model fit compared to a model with only by-subject random intercept ($\chi^2(1) = 0.64$; $p = 0.43$). This supports our assumption that participants did not analyze the meaning of pseudowords.

The model comparison revealed that the best model includes by-subject random slope for quantifier ($\chi^2(2) = 24.76$; $p < 0.001$). The by-subject random slope for response ($\chi^2(2) = 19.90$; $p < 0.001$), and by-subject random slope for proportion ($\chi^2(2) = 12.95$; $p = 002$) improved model fit less. In the next step, we included by-subject random slope for response ($\chi^2(3) = 20.94$; $p = 0001$). The model with three random slopes was overfitted.

B.1.2 Effect of threshold on reaction time – regression model comparison

More than half and *fewer than half*

For *more than half*, we included only by-subject and by-item random intercepts, because a random slope for a response did not improve the model ($\chi^2(2) = 4.24$; $p = 0.12$), and a model with a random slope for distance did not converge. For

fewer than half we included by-subject and by-item random intercepts and by-subject random slope for response ($\chi^2(2) = 14.42$; $p = .001$). The model with by-subject random slope for distance did not converge.

Most

For *most* we included by-subject random slope for response ($\chi^2(2) = 18.05$; $p = .0001$). The model with by-subject random slope for distance did not converge. We did not include by-item random intercept because the model was overfitted and the intercept did not explain any variance.

Many and *few*

For *many* the best random effects structure included by-subject random intercept only. The model with by-subject random slope for response ($\chi^2(2) = 3.37$; $p = .19$) did not improve the model fit and the model with by-subject random slope for distance did not converge. We did not include by-item random intercept because the model was overfitted and the intercept did not explain any variance.

For *few* we included by-subject random slope for response ($\chi^2(2) = 16.55$; $p = 0.0003$). The model with by-subject random slope for distance did not converge. We did not include by-item random intercept because the model was overfitted and the intercept did not explain any variance.

B.1.3 Additional findings

In addition to testing main hypotheses, we also found differences in other DDM parameters between quantifiers. We tested these differences using BMA parameters. We compared only the pairs of parameters that were not constrained between quantifiers.

Firstly, we found that *more than half* had a higher growth rate than *most* (Figure 3.5, $t(71) = -5.08$; $p < .001$; mean difference -259.39). This finding was expected. We found a proportion effect on reaction times for *most*, but not *more than half*. This effect is also reflected in our modeling results in the difference in the growth rate parameter.

Secondly, we also found a difference between positive and negative quantifiers in BMA non-decision time parameters. We found that non-decision time was longer for *fewer than half* than *more than half* ($t(71) = 5.81$; $p < .001$; mean difference 0.03) and longer for *few* than *many* (Figure 3.5, $t(71) = 5.70$; $p < .001$; mean difference 0.03) Furthermore, we found a significant difference between the BMA starting point for *more than half* and *fewer than half* ($t(71) = -2.61$; $p = .01$; mean difference -0.03) and approached significance for *many* and *few* ($t(71) = -1.97$; $p = .053$; mean difference -0.02). The starting point for positive quantifiers

(*more than half*, *many*) was higher than the starting point for corresponding negative quantifiers (*fewer than half*, *few*).

Finally, we also tested the distance between drift rate asymptotes (distance $= V_L - V_U$). The distance was greater for positive quantifiers than negative quantifiers: *more than half* vs. *fewer than half* (Figure 3.5, $t(71) = 11.87$; $p < .001$; mean difference 0.11), *many* vs. *few* (Figure 3.5, $t(71) = 5.51$; $p < .001$; mean difference 0.07). These findings indicate that the DDM can be useful to capture many properties of natural language quantifiers.

B.2 Replication experiment[1]

B.2.1 Methods

Participants

We collected data from 90 subjects. We excluded 26 participants, based on the same criteria as in Experiment 1.

The final sample consisted of 64 participants (41 male), age: $M = 36$, $SD = 9$; range: 23–65. The final sample represented similar educational backgrounds as in Experiment 1: high school graduates (7 subjects), high school graduates who started college (21 subjects), and college graduates (36 subjects). Participants were paid USD 4 for taking part in the experiment. The study was approved by the Ethics Committee of the Faculty of Humanities of the University of Amsterdam.

Exclusion criteria

We used the same exclusion criterion as in Experiment 1. We excluded 6 fast-guesser participants and 4 for participants who did not meet the monotonicity criterion. We also excluded 16 participants who had participated in similar experiments previously.

Design

The same as in Experiment 1.

Procedure

The procedure was almost the same as in Experiment 1. Only the keyboard keys were changed. We replaced the down arrow key with the K key to move to the next screen, and the right/left arrow keys with the J/L keys to respond.

[1]Supplementary materials for "Uncovering the structure of semantic representations using a computational model of decision-making"

Preprocessing reaction time (RT) data

The same as in Experiment 1.

Mixed-effects regression modeling

We used the same modeling strategy as in Experiment 1.

B.2.2 Results

Descriptive statistics

Table B.1 summarizes the mean reaction times and the proportion of true and false in the replication experiment. We found that, on average, participants verified *more than half* slightly faster than *most*. The difference in the proportion of "true" vs. "false" responses was less apparent.

Table B.1: Mean reaction times in seconds (*SD*) and proportion of responses true vs. false in the replication experiment.

Quantifier	Response true		Response false	
	RT	response	RT	response
Few	1.282 (.387)	.40	1.228 (.686)	.60
Fewer than half	1.082 (.109)	.47	1.043 (.156)	.53
Many	1.068 (.275)	.60	1.077 (.174)	.40
Most	.929 (.129)	.49	.987 (.136)	.51
More than half	.914 (.187)	.51	.968 (.219)	.49

Model fit and comparison

We constrained the *Ter* parameter (Model 2) and a parameter (Model 3) to be the same across quantifiers. We also found that the z parameter (Model 4) and V_L, V_U parameters (Model 6) were the same for positive and negative quantifiers. We constrained the s parameter to be the same for *more than half* and *fewer than half* and the same for *few*, *many*, and *most* (Model 5). Next, we found that symmetric boundaries between true and false responses improved fit for some participants (Model 7). Finally, we constrained the p_0 parameter to be zero for *more than half* and *fewer than half* (Model 8) and for *most* (Model 9). Table B.2 summarizes model comparison, Table B.3 shows Bayesian Model Averaged (BMA) parameters, and Figure B.1 presents the fit of Model 7.

Table B.2: Model comparison in the replication experiment (M is Model; k is the number of free parameters in the model; *mean* is the mean rank; n best is the number of participants for whom the given model was the best; n top 3 is the number of participants for whom the given model was one of the three best models).

M	Parameters			Rank		
	Free	Fixed	k	*mean*	n best	n top 3
1	*Ter*, *a*, *z*, p_0, *s*, V_L, V_U		35	6.38	5	13
2	*a*, *z*, p_0, *s*, V_L, V_U	*Ter*	31	6.69	3	7
3	*z*, p_0, *s*, V_L, V_U	*Ter*, *a*	27	5.59	5	11
4	p_0, *s*, V_L, V_U	*Ter*, *a*, *z*	24	5.20	2	13
5	p_0, V_L, V_U	*Ter*, *a*, *z*, *s*	21	4.48	10	23
6	p_0	*Ter*, *a*, *z*, *s*, V_L, V_U	15	4.39	5	29
7	p_0	*Ter*, *a*, *z*, *s*, V_L, V_U	13	4.56	6	27
8		*Ter*, *a*, *z*, *s*, V_L, V_U, p_0	11	**3.45**	12	**36**
9		*Ter*, *a*, *z*, *s*, V_L, V_U, p_0	10	4.25	**16**	33

Table B.3: Summary for model mean (and *SD*) parameters after BMA using AIC, replication experiment.

Quantifier	p_0	*s*	V_L	V_U	*a*	*z*	*Ter*
Few	-.27	160	-.15	.17	.23	.50	.40
	(.34)	(244)	(.06)	(.07)	(.06)	(.05)	(.09)
Fewer than half	-.002	489	-.15	.17	.22	.51	.40
	(.11)	(355)	(.06)	(.05)	(.05)	(.04)	(.09)
Many	-.27	150	-.21	.20	.22	.54	.40
	(.39)	(246)	(.08)	(.06)	(.05)	(.06)	(.09)
Most	.08	154	-.24	.23	.23	.54	.39
	(.17)	(233)	(.08)	(.09)	(.09)	(.06)	(.10)
More than half	-.0001	466	-.22	.22	.23	.53	.39
	(.06)	(349)	(.08)	(.08)	(.05)	(.05)	(.09)

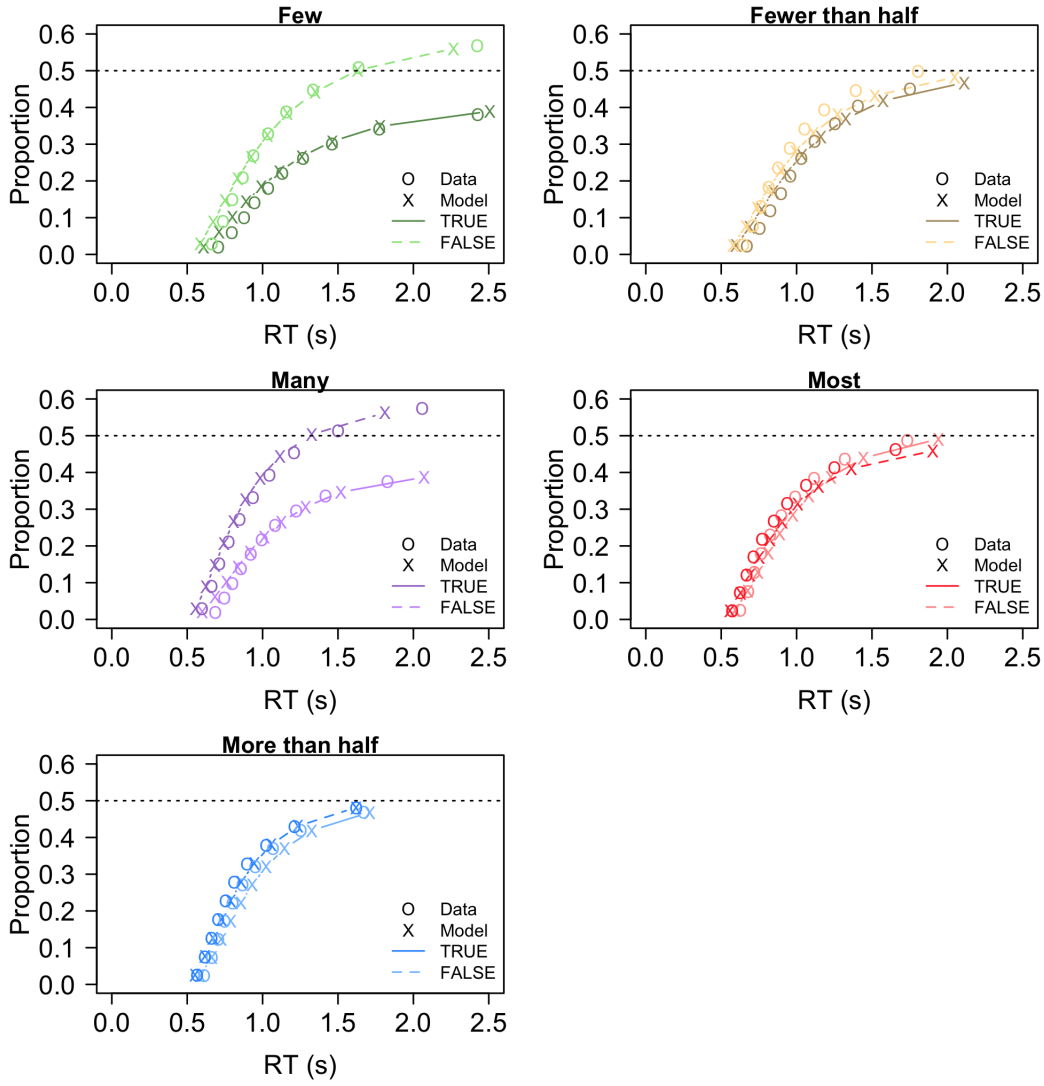


Figure B.1: Defective cumulative density plots show the average fit of Model 7 in the replication experiment. For this visualization, we plot the mean 5%, 15%, 25%, 35%, 45%, 55%, 65%, 75%, 85%, and 95% percentiles over participants, scaled by the proportion of true and false responses, separately for the data and the Model 7 prediction.

Accuracy with respect to threshold

Table B.4 summarizes the mean accurate relative to participants' individual thresholds. The accuracy was slightly lower in the replication experiment. The accuracy for *few* for true responses and *many* for false responses was below 90%.

Table B.4: Relative to threshold mean accuracy (*SD*) in the replication experiment.

Quantifier	Response true	Response false
Few	.88 (.12)	.93 (.12)
Fewer than half	.91 (.08)	.95 (.07)
Many	.95 (.05)	.87 (.19)
Most	.97 (.03)	.96 (.05)
More than half	.97 (.05)	.96 (.05)

LT and ID hypotheses

As in Experiment 1, we tested whether the p_0 parameter could be constrained for *more than half*, *fewer than half*, and *most*. We found that the model with p_0 = 50% for *more than half* and *fewer than half* (Model 8) was better than the Model 7 for 49 out of 64 participants. When we also introduced the constraint $p_0 = 50\%$ for *most* (Model 9), the new model was better than Model 7 for 37 out of 64 participants.

DV hypothesis

As in Experiment 1, we tested the effect of proportion on reaction times for *most* and *more than half*. Firstly, we tested the random effect structure. We included by-subject random intercept and by-subject random slope for response ($\chi^2(2) = 20.93$; $p < .001$). The random slope for proportion improved the model less ($\chi^2(2) = 12.62$; $p = .002$), the random slope for quantifier did not improve the model ($\chi^2(2) = 4.69$; $p = 0.1$), and the model with random slope for response and proportion did not improve model fit ($\chi^2(3) = 7.14$; $p = .07$). We did not include the by-item random intercept because of model overfit. We noted that the variance explained by by-item random intercept was very low.

Table B.5: Summary of the model testing DV hypothesis in the replication experiment.

Effect	Estimate	t value	p value
intercept	-.02	-1.28	.20
prop	-.07	-7.85	$< .001$
quant	-.06	-5.08	$< .001$
resp	.03	2.44	$< .05$
prop:quant	.06	4.74	$< .001$
prop:resp	.14	12.08	$< .001$
quant:resp	.003	.20	.84
prop:quant:resp	-.11	-6.93	$< .001$

prop = proportion; quant = quantifier; resp = response

In the next step, we tested fixed effects of the model. We replicated the effect of proportion (β = -.07; t = -7.85; $p < .001$), the effect of quantifier (β = -.06; t = -5.08; $p < .001$), and quantifier-proportion interaction (β = .06; t = 4.74; $p < .001$). Table B.5 presents the whole model summary. We replicated the finding from Experiment 1 – verification of *most* was proportion-dependent and slower when the proportion was close to 50%.

Effect of threshold on reaction times

Table B.6: Summary of the model estimates testing the effect of threshold, replication experiment.

Effect	*More than half*	*Fewer than half*	*Most*	*Few*	*Many*
intercept	-.08***	-.004	-.02	.11***	.03.
dist	-.0004	-.0003	-.003***	.004***	-.002***
thr	.002	-.003	.0007	-.006**	-.004**
resp	.03**	-.01	.04**	-.002	-.009
dist:thr	-.0002**			-.0001***	
dist:resp	.0009*		.005***	-.006***	.004***
thr:resp				.002	
dist:thr:resp				.0001**	

dist = distance; thr = threshold; resp = response;
*** p < .001; ** p < .01; * p < .05.

More than half and fewer than half For *more than half* and *fewer than half* we did not expect to find a significant effect of threshold. We included by-subject and by-item random intercepts for *more than half*. For *more than half* the model with a random slope for distance did not converge and random slope for response did not improve fit ($\chi^2(2) = 2.29$; $p = .32$). For *fewer than half* we did not include by-item random intercept and by-subject random slopes (for response ($\chi^2(2) = .90$; $p = .64$, for proportion overfit).

For *more than half* we found that the effect of threshold was not significant (β = .002; t = .29; p = .77). However there was significant threshold-distance interaction (β = -.0002; t = -3.01; p = .003). For *fewer than half*, we found that the effect of threshold was not significant (β = -.003; t = -.72; p = .48).

Most For *most* we included by-subject and by-item random intercepts, and by-subject random slope for response ($\chi^2(2) = 23.47$; $p < .001$). The model with by-subject random slope for distance did not converge. We did not replicate the effect of threshold (β = .0007; t = 0.28; p = .78).

Many and few For *many* we did not include the by-item random intercept. We included by-subject random slopes for response ($\chi^2(2) = 20.33$; $p < .001$);

by-subject random slope for distance gave a model that did not converge. We found a significant effect of threshold (β = -.004; t = -3.38; p = .001).

For **few** we included by-subject and by-item random intercepts, and by-subject random slope for response ($\chi^2(2)$ = 11.76; p = .003). The model with by-subject random slope for distance did not converge. We found a significant main effect of threshold (β = -.006; t = -3.29; p = .001) and threshold-distance interaction (β = -.0001; t = -3.83; p ¡ .001).

Although the results of Experiment 1 and its replication are not completely consistent, we found similar patterns. The effect of threshold was present in vague quantifiers like *many* and *few*, but not in quantifiers with sharp meaning boundaries like *more than half* and *fewer than half*. We did not, however, replicate the effect of threshold for *most*. Table B.6 summarizes the estimates for all models.

Additional findings

As in Experiment 1, we also tested the difference in growth rate between *most* and *more than half*. We used BMA parameters. We found that the growth rate was higher for *most* than for *more than half* ($t(63)$ = -6.13; $p < .001$; mean difference -312.01).

We also found differences between positive and negative quantifiers. The BMA starting point was higher for *more than half* than *fewer than half* ($t(63)$ = -2.77; $p < .01$; mean difference -0.03) and for *many* than for *few* ($t(63)$ = -3.89; $p <$ .001; mean difference -0.04). The BMA distance between drift rate asymptotes was greater for *more than half* than *fewer than half* ($t(63)$ = 7.68; $p < .001$; mean difference 0.12) and for *many* than for *few* ($t(63)$ = 5.69; $p < .001$; mean difference 0.08).

In Experiment 1 we constrained the non-decision time parameter to be the same between positive and negative quantifiers. In the replication experiment this constraint did not improve the model. We tested the differences in the *Ter* parameter in Model 1. We did not find significant difference between *more than half* and *fewer than half* ($t(63)$ = 1.35; p = .18; mean difference 0.03), and for *many* and for *few* ($t(63)$ = -.68; p = .50; mean difference -0.01).

Appendix C

Appendix to Chapter 5

C.1 Reaction time analysis without timeout

We conducted an additional analysis of reaction times after excluding the timeout reaction times. Figure C.1 shows that the timeout reaction times were much longer than other reaction times. This affected the shape of the reaction time distribution. The timeout reaction times constituted around 20% of all reaction times. There were more timeout reaction times for *fewer than half* (around 24% per sentence type) than *more than half* (17% and 12% for false and true sentences, respectively).

In the mixed-effects model we included by-subject random slope for trial ($\chi^2(1) = 117.56, p < 0.001$). We found that the interaction was insignificant ($\beta = -0.03, t = -0.61, p = 0.54$), and therefore, we excluded it from the model ($\chi^2(1) = 0.37, p = 0.54$). The main effects of the models without interaction were significant: Quantifier ($\beta = -0.08, t = -2.79, p < 0.005$) and Truth value ($\beta = -0.07, t = -2.42, p < 0.02$), meaning that *fewer than half* was verified slower than *more than half* and that the false sentences were verified slower than true sentences.

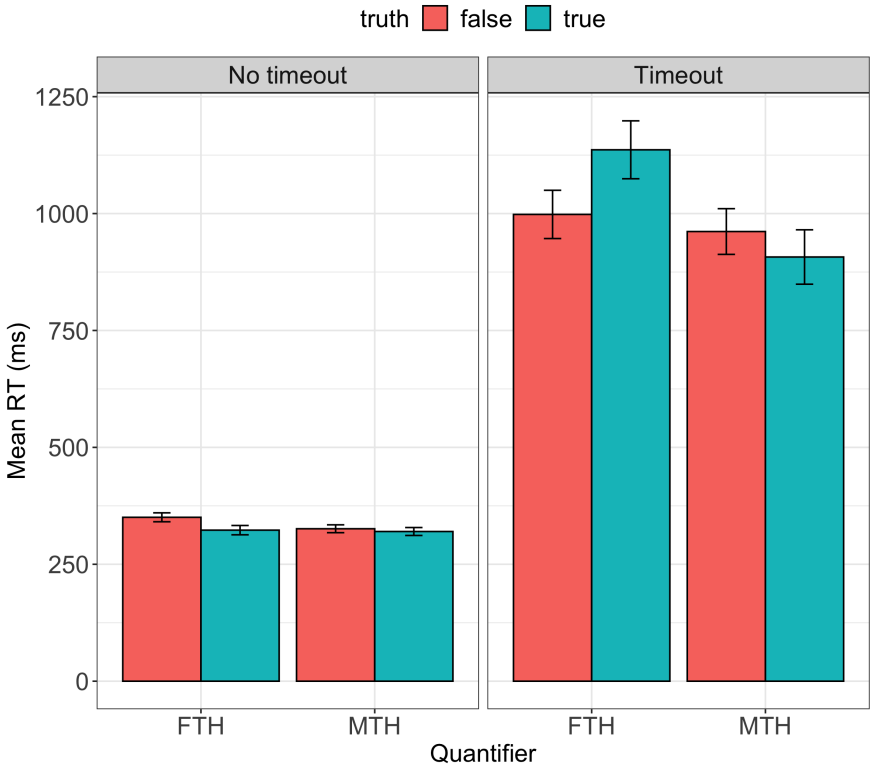


Figure C.1: Comparison between mean reaction times without timeout and timeout reaction times for short sentences. *Fewer than half* is abbreviated as FTH and *more than half* as MTH. The error bars represent within-participant SE.

C.2 ERP analyses

C.2.1 ERPs after the quantifier onset - ANOVA

In addition to analysis in the whole 800 ms time window, we also performed the ERP analysis in the 300–400 ms and 450–800 ms time windows on the averaged EEG signal from the four regions of interest. We ran a two-way repeated measure ANOVA (R function, *anova test*, Quantifier (two levels: *more than half*, *fewer than half*) x ROI (four levels: ROI1, ROI2, ROI3, ROI4)) with the Greenhouse-Geisser correction automatically applied if the within-subject factors violated the sphericity assumption.

In the first time window (300–400 ms), we found significant effect of ROI ($F(1.95,38.99) = 39.62$; $p < 0.001$) and insignificant effects of Quantifier ($F(1,20) = 0.32$; $p = 0.58$), and interaction ($F(1.71,34.19) = 2.23$; $p = 0.09$).

In the second time window (450–800 ms), we found a significant effect of ROI ($F(1,20) = 8.53$; $p = 0.009$) and a significant effect of Quantifier ($F(1.90,38) = 35.04$; $p < 0.001$) and a significant interaction ($F(1.51,30.17) = 6.42$; $p < 0.001$).

This finding replicates the Augurzky et al. (2020) finding. The effect of quantifier was only present for the second time window.

C.2.2 ERPs after the adjective onset - ANOVA

In order to replicate the Augurzky et al. (2020) findings after the adjective onset, we ran a three-way repeated measure ANOVA (R function, *anova test*) with three factors Quantifier (two levels: *more than half*, *fewer than hlaf*), Truth value (two levels: *sentence true*, *sentence false*) and ROI (4 levels: ROI1, ROI2, ROI3, ROI4). First, we computed averaged ERPs from electrodes in each region of interest and then we averaged the signal in the two time windows: 300–400 ms and 450–800 ms after the adjective onset. The Greenhouse-Geisser correction was automatically applied if the within-subject factors violated the sphericity assumption.

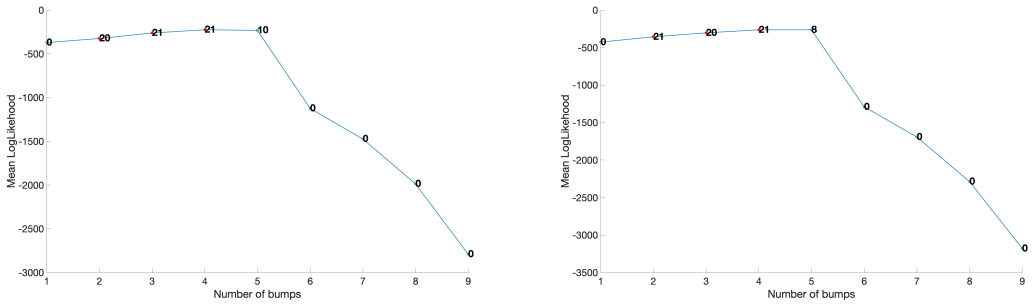
In the first time window (300–400 ms), we found significant effects of Truth value ($F(1,20) = 18.73$, $p = 0.0003$) and ROI ($F(1.75,35.03) = 11.20$, $p = 0.0003$). More importantly, we found a significant interaction between Quantifier and Truth value ($F(1,20) = 14.73$, $p = 0.001$), and an interaction between Truth value and ROI ($F(2.06,41.20) = 3.67$, $p = 0.03$). The analysis of the Quantifier and Truth value interaction revealed that the Truth value effect was present only for *more than half* in all regions of interest: ROI1: $F(1,20) = 22.5$, $p = 0.0001$; ROI2: $F(1,20) = 22.1$, $p = 0.0001$; ROI3: $F(1,20) = 33.7$, $p = 0.0001$; ROI4: $F(1,20) = 28.9$, $p = 0.00003$).

In the second time window (450–800 ms), we found a significant effect of ROI ($F(3,60) = 5.82$, $p = 0.001$), a significant Truth value x Quantifier interaction ($F(1,20) = 6.62$, $p = 0.02$), and three-way interaction ($F(3,60) = 7.47$, $p = 0.0002$). We followed the significant three-way interaction by further analysis of the simple two-way interaction between Quantifier and Truth value. The simple two-way interaction was significant in all regions of interest. Moreover, in the regions ROI1: $F(1,20) = 10.1$, $p = 0.005$; ROI3: $F(1,20) = 10.1$, $p = 0.02$; and ROI4: $F(1,20) = 7.12$, $p = 0.02$, we found a significant Truth value simple effect for *more than half*, but not for *fewer than half*.

In conclusion, these findings are compatible with the cluster-based permutation analyses and with the Augurzky et al. (2020) findings for long sentences. The N400 was present for *more than half*, but not *fewer than half* in the 300 to 400 ms time window after the adjective onset. In addition, the *more than half* conditions differed in the later time window.

C.3 HsMM-MVPA - 500 ms time window after the quantifier onset

The time window 800 ms after the quantifier onset also included the onset of the next words of the sentences (German *als*), which appeared after 500 ms. This overlap could cause a potential confound because some of the stages of processing of the first word could have still continued when the second word was displayed. Therefore, we conducted additional HsMM-MVPA in the time window 500 ms after the quantifier onset, which included only the time when participants saw the first word of the sentence on the screen. For this analysis, we applied the same preprocessing steps already described in the Methods section. We excluded outliers based on reaction times, incorrect responses, and incomplete trials with missing EEG data due to artifacts. The 10 PCA components selected for HsMM-MVPA explained 93.80% of variance.



(a) (b)

Figure C.2: The mean log-likelihood for C.2a *fewer than half* and C.2b *more than half* in 500 ms time window after the quantifier onset. The values indicate for how many participants of 21 included in analysis the more complex model (with n+1 bump) was better than the previous simpler model (with n bumps). The significant increase is indicated by the red point (sign test $p < 0.05$). The plot shows an increase in mean log-likelihood until the 5-bump model, but the 4-bump model is the best model.

We ran two separate HsMM-MVPA models for each condition. In contrast to the analysis in the 800 ms time window, there was no ambiguity in the best model fit for 500 ms time window. The 4-bump models fitted the best in both conditions. The model fit improved for all 21 subjects compared to 3-bump model. The mean log-likelihood of the winning model was LL= -225.01 for *fewer than half* and LL = -263.225 for *more than half*. The 5-bump model with a slightly higher mean log-likelihood improved the model fit for only 10 participants for *fewer than half* and 8 participants for *more than half*. We concluded that there was no evidence for an extra processing step in the first 500 ms time window after the quantifier

onset. Figure C.2 presents the mean log-likelihoods for each model and Figure C.3 presents the stage durations and bumps onsets .

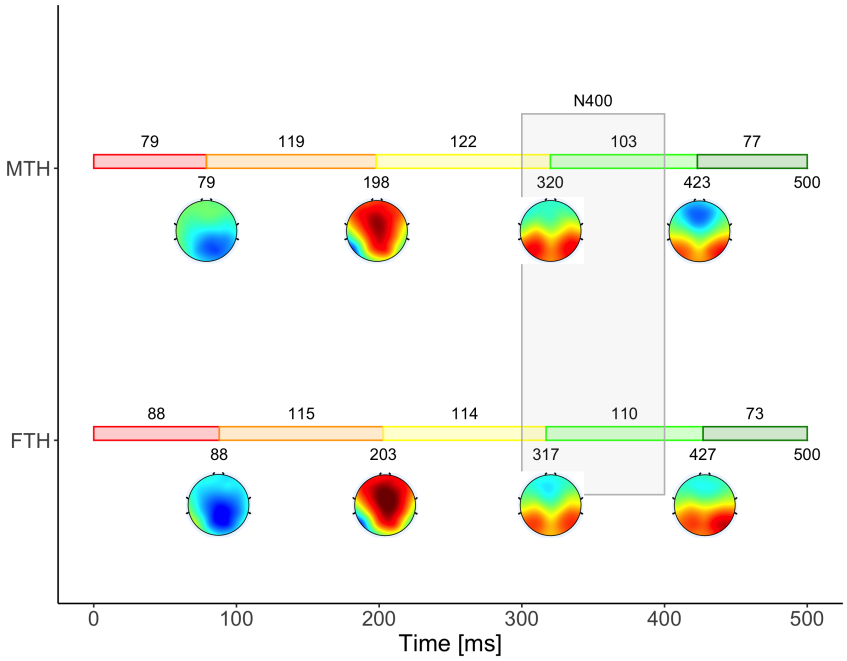
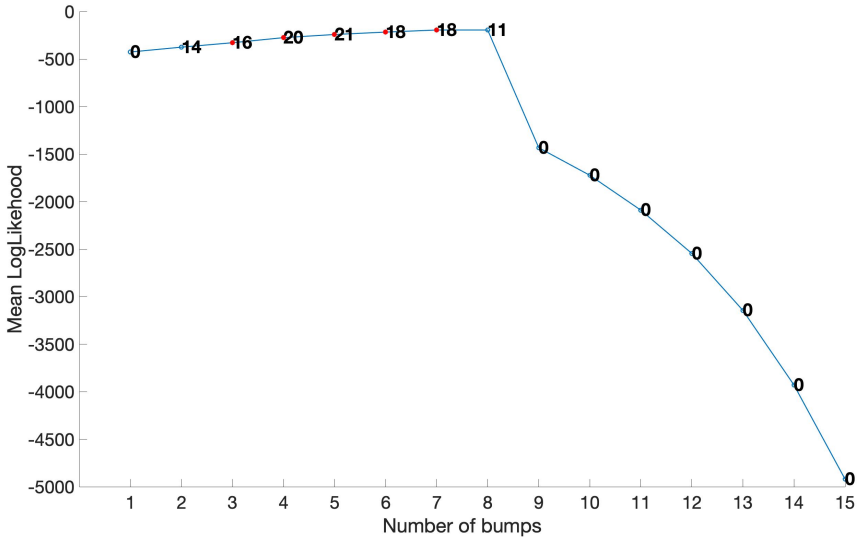


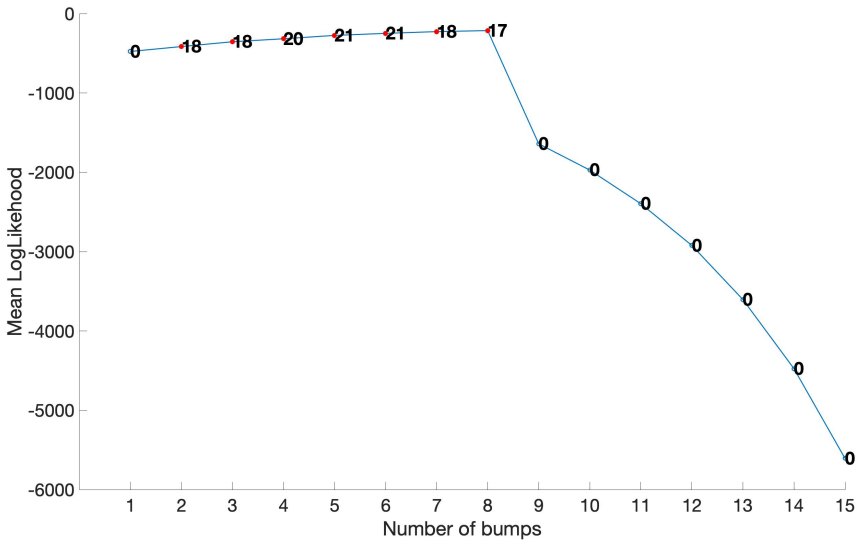
Figure C.3: Bump topographies and stage durations for *more than half* (MTH) and *fewer than half* (FTH) in time window 500 ms after the quantifier onset. The values above bump topographies correspond to the average onset of the bump. The colored bars indicate the stage durations. The values above the colored bars show the mean stage durations. Additionally, the gray lines indicate the ERP analysis time windows from Augurzky et al. (2020).

C.4 HsMM-MVPA - mean log-likelihood

C.4.1 After the quantifier onset



(a)



(b)

Figure C.4: The mean log-likelihood for C.4a *fewer than half* and C.4b *more than half*. The values indicate for how many participants of 21 included in analysis the more complex model (with n+1 bump) was better than the previous simpler model (with n bumps). The significant increase is indicated by red point (sign test $p < 0.05$). C.4a shows the increase in mean log-likelihood until the 8-bump model. However, this model was not better than the simpler model for a significant number of participants. C.4b indicates the increase in mean log-likelihood until the 8-bump model for a significant number of participants.

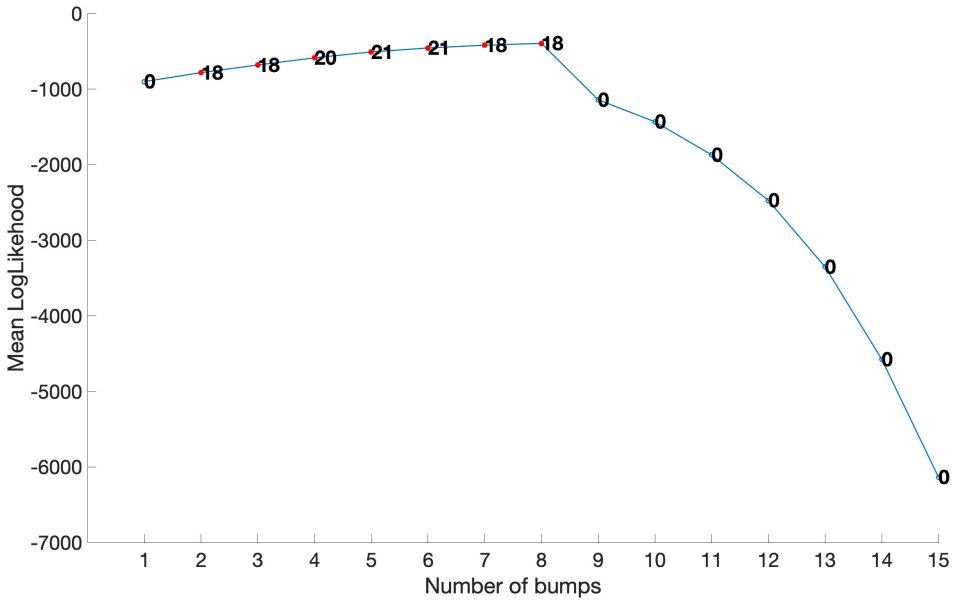


Figure C.5: The mean log-likelihood for the combined model after the quantifier onset. The values indicate for how many of the 21 participants included in the analysis the more complex model (with n+1 bump) was better than the previous simpler model (with n bumps). The significant increase is indicated by the red point (sign test $p < 0.05$). The plot shows the increase in mean log-likelihood until the 8-bump model.

C.4.2 After the adjective onset

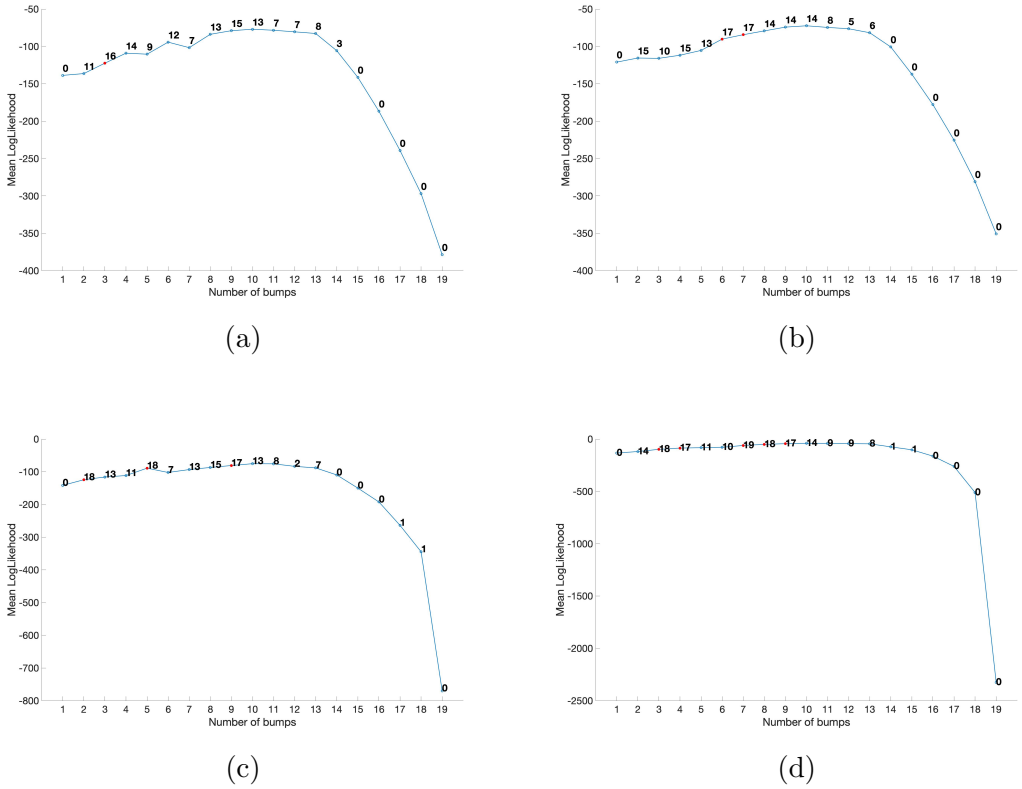


Figure C.6: The mean log-likelihood for C.6a *fewer than half* false sentence, C.6b *fewer than half* true sentence, C.6c *more than half* false sentence, C.6d *more than half* true sentence. The values indicate for how many of the 21 participants included in the analysis the more complex model (with n+1 bump) was better than the previous simpler model (with n bumps). The significant increase is indicated by the red point (sign test $p < 0.05$). For all conditions the mean log-likelihood increased until the 10-bump model. However, none of these models was better than the simpler model for a significant number of participants.

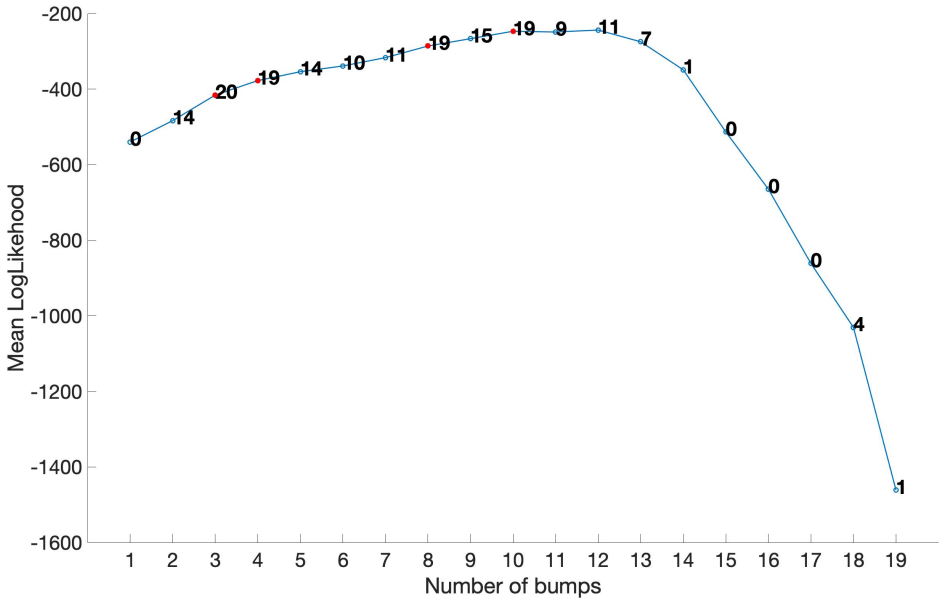


Figure C.7: The mean log-likelihood for the combined model after the adjective onset. The values indicate for how many of the 21 participants included in the analysis the more complex model (with n+1 bump) was better than the previous simpler model (with n bumps). The significant increase is indicated by the red point (sign test $p < 0.05$). The plot shows the increase in mean log-likelihood until the 10-bump model.

C.5 Model comparison

C.5.1 After the quantifier onset

Table C.1 presents the model comparison after the quantifier onset, including the combined model (abbreviation Co), separate models (abbreviation Sep), and mapped models with varied stage duration (Sn), and bump (abbreviations Bn). In addition, we tested models in which Stages 4, 6, 7, and 8, and preceding them, Bumps 3, 5, 6, and 7, varied between conditions (abbreviations S4b3, S6b5, S7b6, S8b7). The numbers in Tables C.1 indicate for how many participants the model in a row was better than the model in a column (sign test based on mean log-likelihood, $p < 0.05$).

Table C.1: Model comparison after the quantifier onset, part I.

	Co	Sep	St1	St2	St3	St4	St5	St6	St7	St8
Co	0	15	9	13	14	14	14	14	14	14
Sep	6	0	6	6	6	6	6	6	6	6
St1	12	15	0	9	13	13	13	13	13	13
St2	8	15	12	0	10	10	10	10	10	10
St3	7	15	8	11	0	0	0	0	0	0
St4	7	15	8	11	0	0	0	0	0	0
St5	7	15	8	11	0	0	0	0	0	0
St6	7	15	8	11	0	0	0	0	0	0
St7	7	15	8	11	0	0	0	0	0	0
St8	7	15	8	11	0	0	0	0	0	0
B1	10	16	10	11	10	10	10	10	10	10
B2	9	14	9	9	9	9	9	9	9	9
B3	5	14	5	5	5	5	5	5	5	5
B4	6	15	6	8	6	6	6	6	6	6
B5	5	15	5	6	5	5	5	5	5	5
B6	4	15	5	6	4	4	4	4	4	4
B7	8	15	10	10	9	9	9	9	9	9
B8	9	15	9	10	9	9	9	9	9	9
St4b3	5	14	5	5	5	5	5	5	5	5
St6b5	5	15	5	6	5	5	5	5	5	5
St7b6	4	15	5	6	4	4	4	4	4	4
St8b7	8	15	10	10	9	9	9	9	9	9

Table C.1: Model comparison after the quantifier onset, part II, (the * indicates a significant result).

	B1	B2	B3	B4	B5	B6	B7	B8	St4b3	St6b5	St7b6	St8b7
Co	11	12	16	15	16	17*	13	12	16	16	17*	13
Sep	5	7	7	6	6	6	6	6	7	6	6	6
St1	11	12	16	15	16	16	11	12	16	16	16	11
St2	10	12	16	13	15	15	11	11	16	15	15	11
St3	11	12	16	15	16	17*	12	12	16	16	17*	12
St4	11	12	16	15	16	17*	12	12	16	16	17*	12
St5	11	12	16	15	16	17*	12	12	16	16	17*	12
St6	11	12	16	15	16	17*	12	12	16	16	17*	12
St7	11	12	16	15	16	17*	12	12	16	16	17*	12
St8	11	12	16	15	16	17*	12	12	16	16	17*	12
B1	0	13	12	12	16	12	12	13	12	16	12	12
B2	8	0	13	14	11	12	10	9	13	11	12	10
B3	9	8	0	8	11	8	9	7	0	11	8	9
B4	9	7	13	0	12	10	10	9	13	12	10	10
B5	5	10	10	9	0	6	8	6	10	0	6	8
B6	9	9	13	11	15	0	8	9	13	15	0	8
B7	9	11	12	11	15	13	0	7	12	13	13	0
B8	8	12	14	12	15	12	14	0	14	15	12	14
St4b3	9	8	0	8	11	8	9	7	0	11	8	9
St6b5	5	10	10	9	0	6	8	6	10	0	6	8
St7b6	9	9	13	11	15	0	8	9	13	15	0	8
St8b7	9	11	12	11	13	13	0	7	12	13	13	0

C.5.2 After the adjective onset

Table C.2 presents the model comparison after the adjective onset, including the combined model (abbreviation Co), separate models (abbreviation Sep), and mapped models with varied stage duration (Sn), and bump (abbreviations Bn). In addition, we tested the following models: Stages 4 and 5 differed between *more than half* false sentences and were the same for other conditions (abbreviations S4mf, S5mf); Stage 7 differed between *fewer than half* false sentences, *more than half* true sentences, and was the same for other conditions (abbreviation S7fm); Stage 9 differed for *more than half* true sentences and was the same for other conditions (abbreviation S9mt); and Stage 11 differed between *fewer than half* and *more than half* (abbreviation S11fm); Stage 11 differed between *more than half* true sentences, *more than half* false sentences, and was the same for *fewer than half* true and false sentences (abbreviation S11fmm); the Bump 4 was different for *more than half* true sentences and *more than half* false sentences and the same for *fewer than half* true and false sentences (abbreviation B4fmm).

The numbers in Table C.2 indicate for how many participants the model in a row was better than the model in a column (sign test based on mean log-likelihood, $p < 0.05$).

Table C.2: Model comparison after the adjective onset, part I, (the * indicates a significant result).

	Co	Sep	St1	St2	St3	St4	St5	St6	St7	St8	St9	St10	St11
Co	0	15	13	13	13	17*	17*	13	16	13	16	13	8
Sep	6	0	5	5	5	6	6	5	6	5	5	5	4
St1	8	16	0	0	0	19*	20*	0	13	0	15	0	8
St2	8	16	0	0	0	19*	20	0	13	0	15	0	8
St3	8	16	0	0	0	19*	20*	0	13	0	15	0	8
St4	4	15	2	2	2	0	8	2	3	2	2	2	4
St5	4	15	1	1	1	13	0	1	2	1	1	1	6
St6	8	16	0	0	0	19*	20*	0	13	0	15	0	8
St7	5	15	8	8	8	18	19*	8	0	8	14	8	6
St8	8	16	0	0	0	19*	20*	0	13	0	15	0	8
St9	5	16	6	6	6	19*	20*	6	7	6	0	6	6
St10	8	16	0	0	0	19*	20*	0	13	0	15	0	8
St11	13	17*	13	13	13	17*	15	13	15	13	15	13	0
S4mf	4	15	2	2	2	13	10	2	3	2	2	2	4
S5mf	4	15	1	1	1	13	0	1	2	1	1	1	6
S7fm	5	15	8	8	8	18*	19*	8	11	8	15	8	6
S9mt	8	16	11	11	11	18*	20*	11	11	11	13	11	7
S11fm	13	16	12	12	12	18*	17*	12	14	12	16	12	9
B1	8	15	7	7	7	17*	18*	7	9	7	12	7	7
B2	5	15	4	4	4	16	14	4	8	4	8	4	6
B3	7	15	7	7	7	17*	15	7	10	7	9	7	7
B4	4	14	2	3	3	10	7	2	5	3	4	2	3
B5	6	15	4	4	4	11	9	4	7	4	7	4	6
B6	7	15	5	5	5	17*	14	5	8	5	10	5	7
B7	7	15	5	5	5	15	15	5	8	5	7	5	6
B8	6	15	6	6	6	15	13	6	7	6	8	6	5
B9	4	15	3	3	3	15	9	3	4	3	3	3	5
B10	6	15	5	5	5	13	13	5	8	5	8	5	6
S11fmm	13	17*	13	13	13	18*	16	13	15	13	15	13	11
B4fmm	4	14	2	2	2	11	9	2	4	4	4	2	3

Table C.2: Model comparison after the adjective onset, part II, (the * indicates a significant result).

	S4mf	S5mf	S7fm	S9mt	S11fm	B1	B2	B3	B4	B5	B6
Comb	17*	17*	16	13	8	13	16	14	17*	15	14
Sep	6	6	6	5	5	6	6	6	7	6	6
St1	19*	20*	13	10	9	14	17*	14	19*	17*	16
St2	19*	20*	13	10	9	14	17*	14	19*	17*	16
St3	19*	20*	13	10	9	14	17*	14	19*	17*	16
St4	8	8	3	3	3	4	5	4	11	10	4
St5	11	0	2	1	4	3	7	6	14	12	7
St6	19*	20*	13	10	9	14	17*	14	19*	17*	16
St7	18*	19*	10	10	7	12	13	11	16	14	13
St8	19*	20*	13	10	9	14	17*	14	19*	17*	16
St9	19*	20*	6	8	5	9	13	12	17*	14	11
St10	19*	20*	13	10	9	14	17*	14	19*	17*	16
St11	17*	15	15	14	12	14	15	14	18*	15	14
S4mf	0	10	3	3	3	5	4	4	14	11	5
S5mf	11	0	2	1	4	3	7	6	14	12	7
S7fm	18*	19*	0	10	7	12	13	11	16	15	13
S9mt	18*	20*	11	0	6	12	15	14	18*	15	15
S11fm	18*	17*	14	15	0	14	17*	16	17*	16	14
B1	16	18*	9	9	7	0	16	11	15	15	12
B2	17*	14	8	6	4	5	0	12	17*	14	10
B3	17*	15	10	7	5	10	9	0	18*	12	11
B4	7	7	5	3	4	6	4	3	0	9	4
B5	10	9	6	6	5	6	7	9	12	0	9
B6	16	14	8	6	7	9	11	10	17*	12	0
B7	14	15	8	6	6	8	9	10	14	13	9
B8	14	13	7	6	5	6	10	9	13	11	7
B9	13	9	5	3	4	5	7	5	15	9	6
B10	12	13	8	6	6	6	11	6	16	12	7
S11fmm	18*	16	15	14	11	14	15	14	18*	16	14
B4fmm	9	9	4	4	4	6	5	4	17*	8	4

Table C.2: Model comparison after the adjective onset, part III, (the * indicates a significant result).

	B7	B8	B9	B10	S11fmm	B4fmm
Comb	14	15	17	15	8	17
Sep	6	6	6	6	4	7
St1	16	15	18*	16	8	19*
St2	16	15	18*	16	8	19*
St3	16	15	18*	16	8	19*
St4	6	6	6	8	3	10
St5	6	8	12	8	5	12
St6	16	15	18*	16	8	19*
St7	13	14	17*	13	6	17*
St8	16	15	18*	16	8	19*
St9	14	13	18*	13	6	17*
St10	16	15	18*	16	8	19*
St11	15	16	16	15	10	18*
S4mf	7	7	8	9	3	12
S5mf	6	8	12	8	5	12
S7fm	13	14	16	13	6	17*
S9mt	15	15	18*	15	7	17*
S11fm	15	16	17*	15	10	17*
B1	13	15	16	15	7	15
B2	12	11	14	10	6	16
B3	11	12	16	15	7	17*
B4	7	8	6	5	3	4
B5	8	10	12	9	5	13
B6	12	14	15	14	7	17*
B7	0	11	14	10	6	12
B8	10	0	14	10	5	13
B9	7	7	0	9	5	14
B10	11	11	12	0	5	15
S11fmm	15	16	16	16	0	18*
B4fmm	9	8	7	6	3	0

C.6 HsMM-MVPA of long sentences

We wanted to test which stages are related to sentence processing and which are specific for truth value evaluation and response. We contrasted short sentences with long sentences. Because long sentences had continuation, we constrained the time window to 900 ms, which was before the onset of the comma indicating sentence continuation. We fitted four separate models to four conditions of inter-

mediate truth evaluation[1]: *fewer than half* false, *fewer than half* true, *more than half* false, and *more than half* true. We followed the same pre-processing steps, downsampled the data to 100 Hz, excluded outliers trials, and ran PCA, as for short sentences. The 10 PCA components selected for HsMM-MVPA explained 92.32% of variance.

Figure C.8 presents the model comparison after leave-one-out cross validation for separate models for four conditions in long sentences. For all conditions the best model had 9 bumps. Figure C.9 presents bumps topographies and stage durations for separate models.

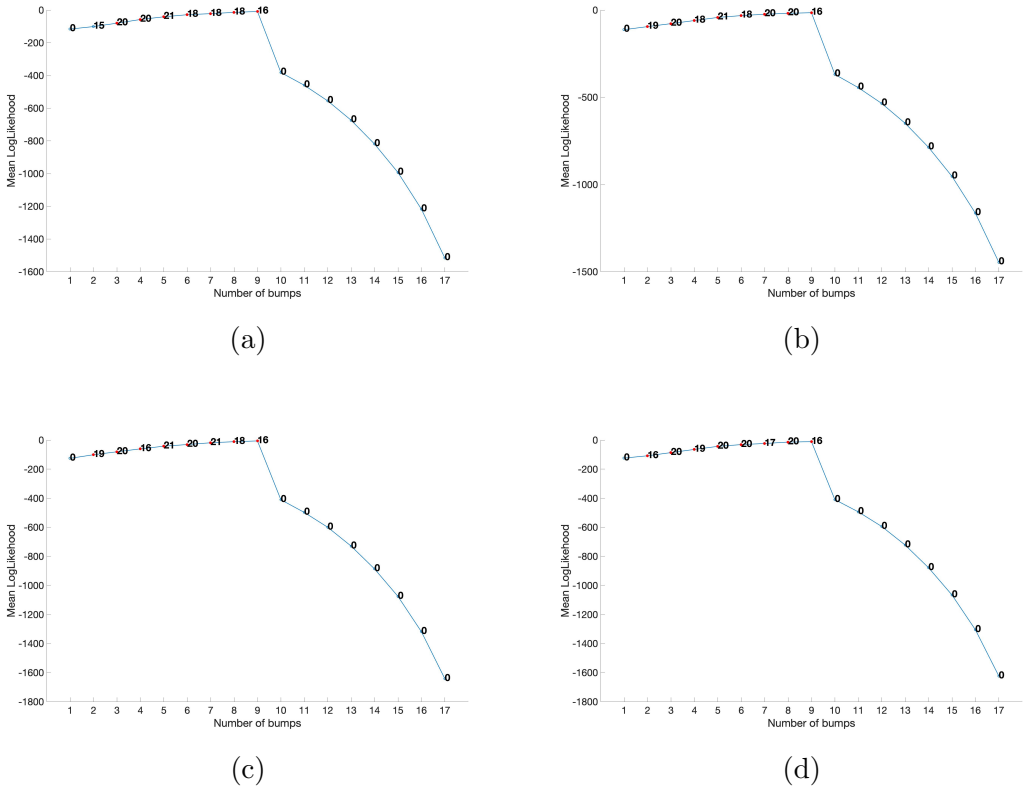


Figure C.8: The mean log-likelihood for the long sentences: C.8a *fewer than half* false, C.6b *fewer than half* true, C.8c *more than half* false, and C.8d. *more than half* true. The values indicate for how many of the 21 participants included in the analysis the more complex model (with n+1 bump) was better than the previous simpler model (with n bumps). The significant increase is indicated by the red point (sign test $p < 0.05$). For all conditions the mean log-likelihood increases in until the 9-bump models.

[1]In the long sentences, the truth value of the sentence could change after participants read the second part of the sentence. Here we refer only to the truth value of the first part of the sentence e.g., "*More than half* of the dots are blue."

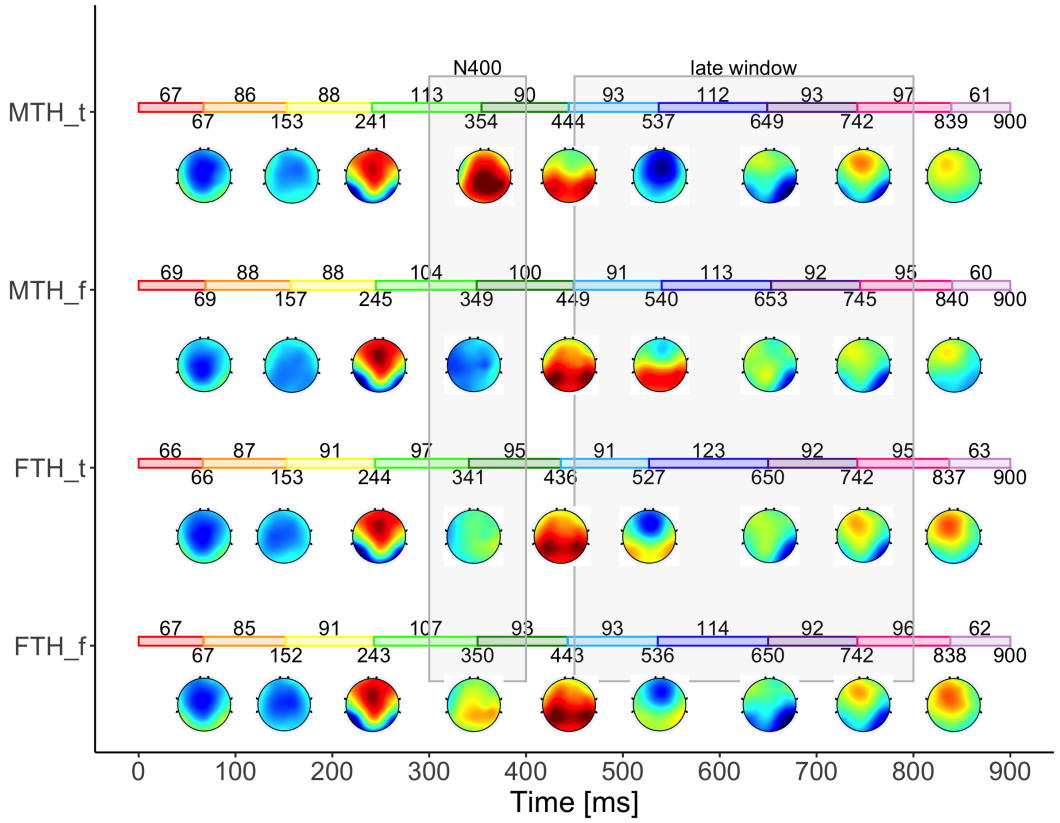


Figure C.9: Bump topographies and stage durations for separate conditions (*more than half* true sentence (MTH_t), *more than half* false sentence (MTH_f), *fewer than half* true sentence (FTH_t), and *fewer than half* false sentence (FTH_f)), long sentences. The values above bump topographies correspond to the average onset of the bump. The colored bars indicate the stage durations. The values above the colored bars show the mean stage durations. Additionally, the gray lines indicate the ERP analysis time windows from Augurzky et al. (2020).

In the next step, we fitted the combined model to test whether, as for short sentences, the combined model would be better than separate models. The combined model 9-bump solution had the best fit (LL = -29.26), for 18 out of 21 participants (sign test $p < 0.05$), meaning that the 9-bump solution was robust in both separate and combined models. The combined model was better than separate models (mean LL = -41.93 for four separate models) for 16 out of 21 participants (sign test p ns.), meaning that the separate models did not outperform the combined model. We did not test the mapped models with different stage durations predicting similar negative results as for short sentences. However, we plotted the stage durations of the combined model to compare them between conditions. Figure C.10 shows the stage durations of the combined model split into four conditions.

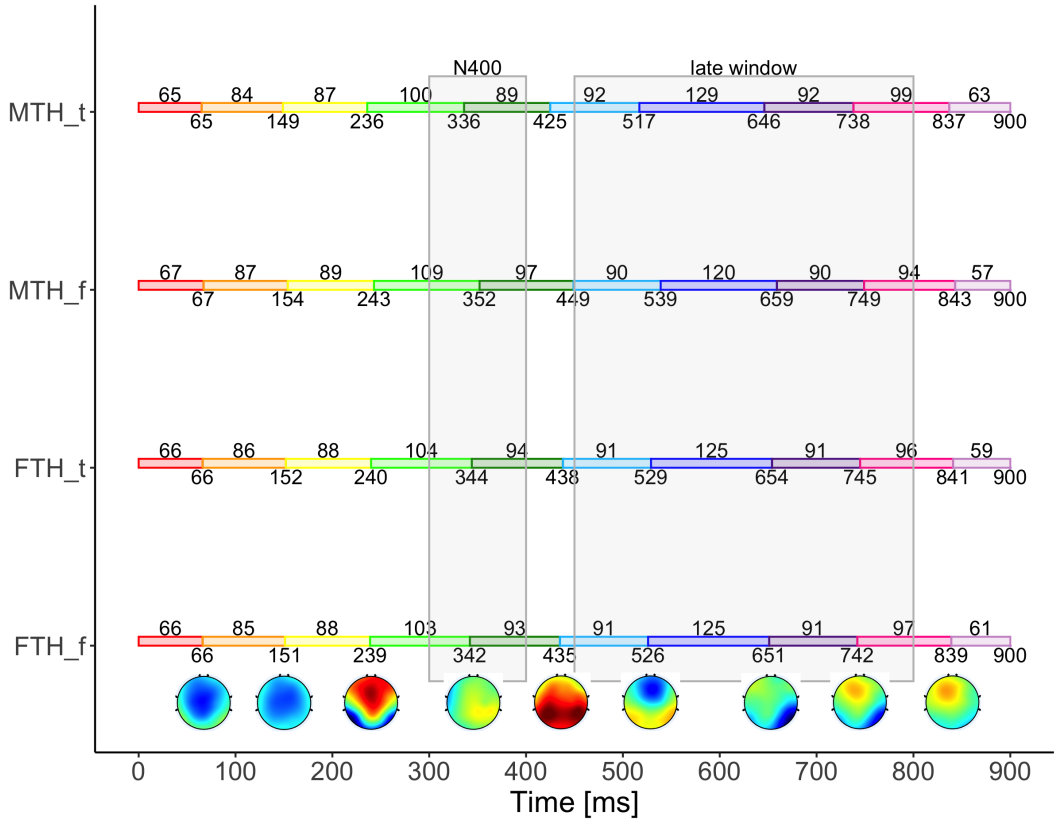


Figure C.10: Bump topographies and stage durations for the combined model (conditions: *more than half* true sentence (MTH_t), *more than half* false sentence (MTH_f), *fewer than half* true sentence (FTH_t), and *fewer than half* false sentence (FTH_f)), long senetnces. The colored bars indicate the stage durations. The values above the colored bars show the mean stage durations and the values below the average onset of the bumps. Additionally, the gray lines indicate the ERP analysis time windows from Augurzky et al. (2020).

In the analysis of short sentences, Stages 7 and 11 differed between conditions. Moreover, Stages 7, 9, and 11 were predictive of the reactions times. Although we could not measure the effect of these stages on reaction times in long sentences, we investigated the differences in Stages 7 and 9 between conditions in long sentences. We visually compared the topographies of bumps preceding these stages with corresponding bumps in short sentences. As in short sentences, Stage 7 in long sentences started with the onset of a frontal negative bump. Stage 9 started with a bump with a left frontal positive activity, comparable to the Bump 8 in short sentences. We concluded that Bumps 6 and 8 in the long sentences correspond to Bumps 6 and 8 in the short sentences. Therefore, we decided to test differences between conditions in the duration of Stages 7 and 9 in the long sentences. We applied the same statistical procedure as in the analysis of short sentences.

C.6.1 Stage 7

We did not include the trial random slope ($\chi^2(1) = 1.49, p = 0.22$) as it did not improve model fit. We found the significant intercept of the model ($\beta = 4.81, t = 330.81, p < 0.001$), main effect of (intermediate) Truth value ($\beta = 0.04, t = 5.03, p < 0.001$), and interaction between (intermediate) Truth value and Quantifier ($\beta = 0.07, t = 4.98, p < 0.001$). The main effect of Quantifier was not significant ($\beta = 0.0008, t = 0.13, p = 0.90$). The finding in Stage 7 in long sentences did not fully replicate the finding in Stage 7 in short sentences. In short sentences, the interaction effect was not present, whereas in long sentences, it was significant. Moreover, in short sentences, there was a significant effect of Quantifier, which was insignificant in long sentences.

C.6.2 Stage 9

In Stage 9, the random slope for trial also did not improve the model fit ($\chi^2(1) = 3.41, p = 0.06$). Similarly to Stage 7, in Stage 9 we also found significant intercept of the model ($\beta = 4.56, t = 480.40, p < 0.001$), main effect of (intermediate) Truth value ($\beta = 0.02, t = 4.00, p < 0.001$), and interaction between (intermediate) Truth value and Quantifier ($\beta = 0.06, t = 6.14, p < 0.001$), but the main effect of Quantifier was not significant ($\beta = 0.001, t = 0.24, p = 0.81$). This finding diverges from the findings in short sentences in which the differences between quantifiers were not significant.

C.7 Stage 11 analysis without timeout

Similar to reaction times distribution, the distribution of Stage 11 also had a somewhat binomial shape. The shape of the distribution deviated from normal even after log-transformation. Therefore, we decided to run an additional analysis on Stage 11 with excluded timeout reaction times.

We included the by-subject random intercept and random slope for trial ($\chi^2(1) = 68.98, p < 0.001$). The Quantifier x Truth value interaction was not significant ($\beta = 0.05, t = 0.90, p = 0.37; \chi^2(1) = 0.80, p = 0.37$). In the model without interaction, only the intercept was significant ($\beta = 5.37, t = 61.21, p < 0.001$), but neither the effect of Quantifier ($\beta = 0.04, t = 1.51, p = 0.13$) nor the effect of Truth value ($\beta = -0.008, t = -0.29, p = 0.77$).

Appendix D

Appendix to Chapter 6

D.1 Bayesian logistic regression model diagnostics

The divergent transitions for all models were zero.

Table D.1: Bayesian logistic regression model diagnostics: *Rhat* values, autocorrelations in chains (autocor.), number of effective samples (n eff), posterior predictive checks (PPchecks).

Model	*Rhat*	autocor.	n eff	PPchecks
monotonicity	1	no	1,397-3,577	good
at least 3 vs. *between 3 and 6*	1	no	1,046-2,618	good
at most 2 vs. *between 3 and 6*	1	chain 3	865-2,871	good
convexity	1	no	1,215-3,188	good
quantity	1	chain 2	1,225-4,264	good
at least 3 vs. *the first 3*	1	chain 2 and 4	1,047-2,448	good
at least 3 vs. *the last 3*	1	no	1,480-2,690	good
conservativity	1	chain 3 and 1	596-2,785	good

D.2 Graphical representation of Bayesian logistic regression model estimates

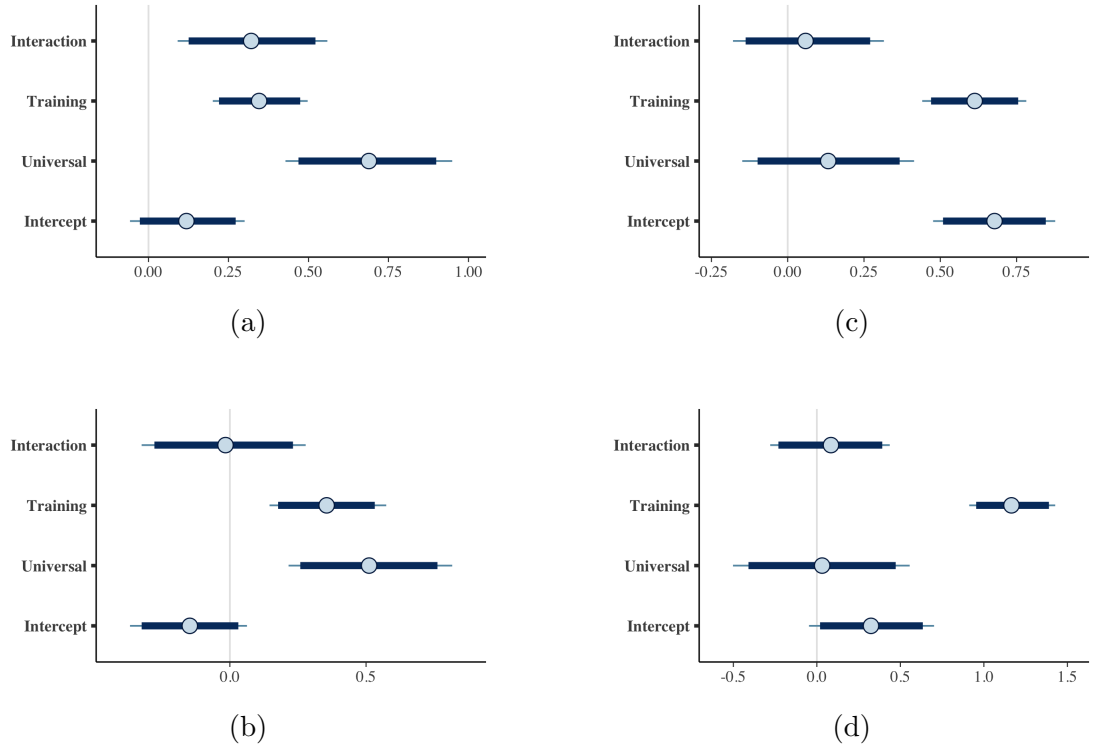


Figure D.1: Posterior distributions for the regression coefficients of each universal: D.1a monotonicity, D.1b convexity, D.1c quantity, D.1d conservativity. The blue dots indicate the median and the bars indicate the 95% credible interval.

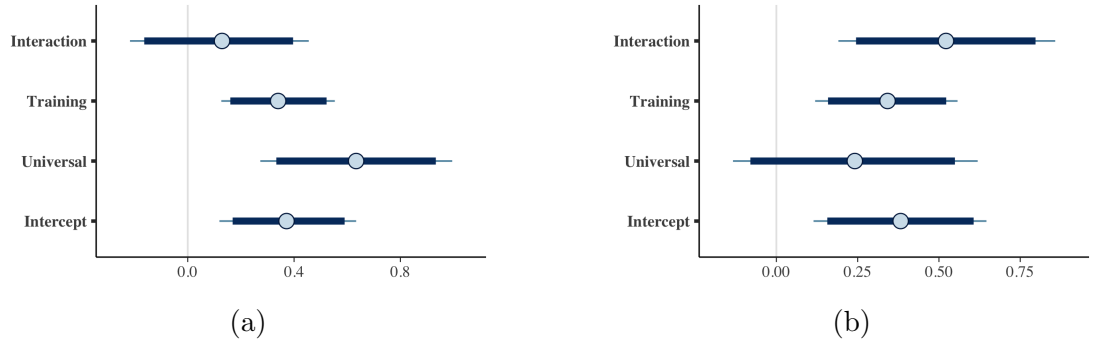
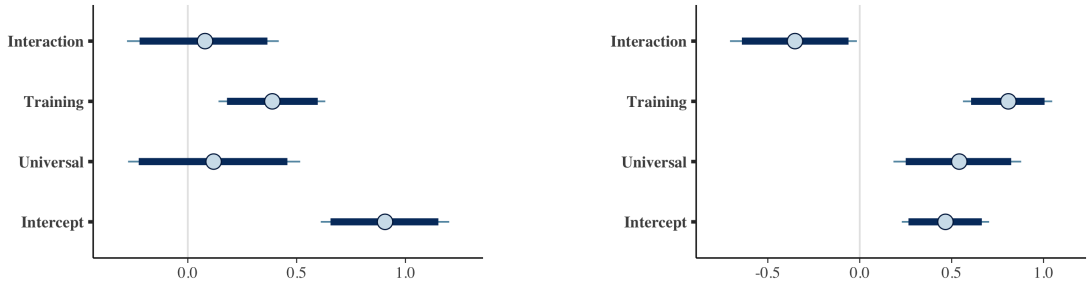


Figure D.2: Posterior distributions for the monotonicity universal: D.2a *at least 3* vs. *between 3 and 6*, D.2b *at most 2* vs. *between 3 and 6*. The blue dots indicate the median and the bars indicate the 95% credible intervals.



(a) (b)

Figure D.3: Posterior distributions for the quantity universal: D.3a *at least 3* vs. *the first 3*, D.3b *at least 3* vs. *the last 3*. The blue dots indicate the median and the bars indicate the 95% credible intervals.

D.3 Models of quantifier frequency

The COCA corpus includes 31,277,351 words tagged as determiners, meaning that they occur with a frequency of 32,233.61 per million words. The most frequent quantifiers in the corpus were *all* (2,505.93 per million words), *some* (1,684.47 per million words), *many* (903.26 per million words), *few* (472.23 per million words), and *most* (93.45 per million words).[1]

Table D.2 summarizes the frequencies of occurrence, the exact quantifier from our study, and the quantifier construction like, DET (NUM) NOUN, where DET is, for example, *at least*, NUM is the number in numerical quantifiers (e.g., 3 or three) and NOUN is the noun which follows the quantifier. We measured the frequency of these quantifiers as a more general measure. The frequency of the exact quantifier is obscured by the choice of the number. However, we think that it should not matter if participants were to learn e.g., *at least 3* or *at least 1*, although the second one is more frequent. We included two standard measures of frequency: frequency per million words (fpmw) and *Zipf* value (van Heuven et al., 2014).

[1]Counts of bare *most*, without *the most*.

Table D.2: Frequency table. For all numerical quantifiers we searched numerical information as a word (e.g., three) or Arabic number (e.g., 3). The columns fpmw show the frequency per million words (and row frequency from the COCA corpus in the brackets) and the columns *Zipf* value show the word frequency computed as log10(fpmw*1,000).

Quantifier	Exact quantifier		Construction	
	fpmw	*Zipf* value	fpmw	*Zipf* value
*at least 3**	17.03 (17053)	4.23	155.96 (156185)	5.19
*at most 2***	2.91 (2910)	3.46	30.65 (30696)	4.49
between 3 and 6	0.10 (102)	2.01	8.90 (8919)	3.95
at most 2 or at least 7	0 (0)	-	0 (0)	-
the first 3	5.03 (5035)	3.70	25.32 (25363)	4.40
the last 3	7.27 (7279)	3.86	54.16 (54246)	4.73
not all	31.06 (31110)	4.49	2.30 (2307)	3.36
not only	129.53 (129717)	5.11	NA	NA

*also *at least three*, *more than 2*, *more than two*.

**also *at most two*, *fewer than 3*, *fewer than three*, *less than 3*, *less than three*, *not more than 2*, *not more than two*.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., ... Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2019). Measuring the cognitive cost of downward monotonicity by controlling for negative polarity. *Glossa: a journal of general linguistics*, *4*(1).

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In K. G. Parzen E. Tanabe K. (Ed.), *Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics).* (pp. 199–213). Springer, New York, NY.

Anders, R., Riès, S., van Maanen, L., & Alario, F. X. (2015). Evidence accumulation as a model for lexical selection. *Cognitive Psychology*, *82*, 57–73.

Anders, R., van Maanen, L., & Alario, F. X. (2019). Multi-factor analysis in language production: Sequential sampling models mimic and extend regression results. *Cognitive Neuropsychology*, *36*(5-6), 234–264.

Anderson, J. R., Borst, J. P., Fincham, J. M., Ghuman, A. S., Tenison, C., & Zhang, Q. (2018). The Common Time Course of Memory Processes Revealed. *Psychological Science*, *29*(9), 1463–1474.

Anderson, J. R., Zhang, Q., Borst, J. P., & Walsh, M. M. (2016). The discovery of processing stages: Extension of Sternberg's method. *Psychological Review*, *123*(5), 481–509.

Ariel, M. (2003). Does most mean 'more than half'? In P. Nowak, G. Yoquelet, & D. Mortensen (Eds.), *Proceedings of the Twenty-Ninth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Phonetic*

Sources of Phonological Patterns: Synchronic and Diachronic Explanations (pp. 17–30). Berkeley, California: Berkeley Linguistics Society.

Augurzky, P., Bott, O., Sternefeld, W., & Ulrich, R. (2017). Are all the triangles blue?-ERP evidence for the incremental processing of German quantifier restriction. *Language and Cognition*, *9*(4), 603–636.

Augurzky, P., Schlotterbeck, F., & Ulrich, R. (2020). Most (but not all) quantifiers are interpreted immediately in visual context. *Language, Cognition and Neuroscience*, *35*(9), 1203–1222.

Bailey, K., Mlynarczyk, G., & West, R. (2016). Slow wave activity related to working memory maintenance in the N-back task. *Journal of Psychophysiology*, *30*(4), 141–154.

Barber, H. A., Otten, L. J., Kousta, S. T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, *125*(1), 47–53.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*(2), 159–219.

Basar, E. (1980). *EEG-Brain dynamics: Relation between EEG and brain evoked potentials*. Amsterdam: Elsevier.

Beltrán, D., Morera, Y., García-Marco, E., & De Vega, M. (2019). Brain inhibitory mechanisms are involved in the processing of sentential negation, regardless of its content. Evidence from EEG theta and beta rhythms. *Frontiers in Psychology*, *10*, 1782.

van Benthem, J. F. (1986). *Essays in Logical Semantics* (Vol. 29). Dordrecht: Springer.

Berberyan, H. S., van Maanen, L., van Rijn, H., & Borst, J. (2020). EEG-based identification of evidence accumulation stages in decision-making. *Journal of Cognitive Neuroscience*, *33*(3), 510–527.

Berberyan, H. S., van Rijn, H., & Borst, J. P. (2021). Discovering the brain stages of lexical decision: Behavioral effects originate from a single neural decision process. *Brain and Cognition*, *153*.

Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision making: Drift-diffusion model is equivalent to a Bayesian model. *Frontiers in Human Neuroscience*, *8*(1), 102.

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., ... Wagenmakers, E. J. (2018). Estimating across-trial variability parameters of the Diffusion Decision Model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *87*, 46–75.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Borst, J. P., & Anderson, J. R. (2015). The discovery of processing stages: Analyzing EEG data with hidden semi-Markov models. *NeuroImage*, *108*(1), 60–73.

Borst, J. P., & Anderson, J. R. (2021). Discovering Cognitive Stages in M/EEG Data to inform Cognitive Models. In B. Forstmann & B. Turner (Eds.), *An introduction to model-based cognitive neuroscience* (2nd ed.). New York: Springer.

Bott, O., Augurzky, P., Sternefeld, W., & Ulrich, R. (2017). Incremental generation of answers during the comprehension of questions with quantifiers. *Cognition*, *166*, 328–343.

Bott, O., Schlotterbeck, F., & Klein, U. (2019). Empty-Set Effects in Quantifier Interpretation. *Journal of Semantics*, *36*(1), 99–163.

Botvinick, M. M., Carter, C. S., Braver, T. S., Barch, D. M., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive science*, *41 Suppl 6*, 1318–1352.

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143.

Carcassi, F., Schouwstra, M., & Kirby, S. (2019). The evolution of adjectival monotonicity. In M. Espinal, E. Castroviejo, M. Leonetti, L. McNally, & C. Real-Puigdollers (Eds.), *Proceedings of Sinn und Bedeutung 23* (Vol. 1, pp. 219–230). Bellaterra (Cerdanyola del Vallès): Universitat Autònoma de Barcelona.

Carcassi, F., Steinert-Threlkeld, S., & Szymanik, J. (2019). The emergence of monotone quantifiers via iterated learning. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (p. 190-196). Montreal, QB: Cognitive Science Society.

Carcassi, F., & Szymanik, J. (2021). 'Most' vs 'More Than Half': An Alternatives Explanation. In A. Ettinger, E. Pavlich, & B. Prickett (Eds.), *Proceedings of the Society for Computation in Linguistics* (Vol. 4, pp. 334–343). Society for Computation in Linguistics.

Carreiras, M., Vergara, M., & Perea, M. (2007). ERP correlates of transposed-letter similarity effects: Are consonants processed differently from vowels? *Neuroscience Letters*, *419*(3), 219–224.

Chater, N., & Oaksford, M. (1999). The Probability Heuristics Model of Syllogistic Reasoning. *Cognitive Psychology*, *38*(2), 191–258.

Chemla, E., Buccola, B., & Dautriche, I. (2019). Connecting Content and Logical Words. *Journal of Semantics*, *36*(3), 531–547.

Chemla, E., Dautriche, I., Buccola, B., & Fagot, J. (2019). Constraints on the lexicons of human languages have cognitive roots present in baboons (Papio papio). *Proceedings of the National Academy of Sciences of the United States of America*, *116*(30), 14926–14930.

Chen, Q., Ye, C., Liang, X., Cao, B., Lei, Y., & Li, H. (2014). Automatic processing of taxonomic and thematic relations in semantic priming — Differentiation by early N400 and late frontal negativity. *Neuropsychologia*, *64*, 54–62.

Clark, H. H. (1976). *Semantics and Comprehension*. The Hague: Mouton & Co.B.V.

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

Clerc, M. (2010). *Particle Swarm Optimization*. Wiley-ISTE.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89–103.

Dautriche, I., & Chemla, E. (2016). What Homophones Say about Words. *PLOS ONE*, *11*(9), e0162176.

Dautriche, I., Chemla, E., & Christophe, A. (2016). Word Learning: Homophony and the Distribution of Learning Exemplars. *Language Learning and Development*, *12*(3), 231–251.

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA).* (Available online at https://www.english-corpora.org/coca/)

Degen, J., & Goodman, N. D. (2014). Lost your marbles? The puzzle of dependent measures in experimental pragmatics. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 397–402). Austin TX: Cognitive Science Society.

Degen, J., & Tanenhaus, M. K. (2019). Constraint-Based Pragmatic Processing. In C. Cummins & N. Katsos (Eds.), *The Oxford Handbook of Experimental Semantics and Pragmatics* (pp. 20–38). Oxford University Press.

Dehaene, S. (1997). *The Number Sense: How the mind creates mathematics.* New York: Oxford University Press.

Dehaene, S. (2007). Symbols and quantities in parietal cortex: elements of a mathematical theory of number representation and manipulation. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Sensorimotor Foundations of Higher Cognition* (pp. 526–574). Oxford University Press.

De Jong, R., Coles, M. G., Logan, G. D., & Gratton, G. (1990). In search of the point of no return: the control of response processes. *Journal of experimental psychology. Human perception and performance*, *16*(1), 164–182.

Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, *135*, 103569.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21.

Delorme, A., Sejnowski, T., & Makeig, S. (2007). Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage*, *34*, 1443–1449.

Denić, M., & Szymanik, J. (2020). Are most and more than half truth-conditionally equivalent?*. , 1–22.

Deschamps, I., Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition*, *143*, 115–28.

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*(C), 412–431.

Donkin, C., Heathcote, A., Brown, S., & Andrews, A. (2009). Non-decision time effects in the lexical decision task. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2902–2907). Austin, TX: Cognitive Science Society.

Donzallaz, M., Haaf, J. M., & Stevenson, C. (2021). Creative or Not? Hierarchical Diffusion Modeling of the Creative Evaluation Process [Unpublished manuscript].

Dudschig, C., & Kaup, B. (2018). How does "Not Left" Become "Right"? Electrophysiological evidence for a dynamic conflict-bound negation processing account. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(5), 716–728.

Dutilh, G., Krypotos, A.-M., & Wagenmakers, E.-J. (2011). Task-Related Versus Stimulus-Specific Practice A Diffusion Model Account. *Experimental Psychology*, *58*, 434–442.

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E. J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin and Review*, *16*(6), 1026–1036.

Farshchi, S., Andersson, A., van de Weijer, J., & Paradis, C. (2020). Processing sentences with sentential and prefixal negation: an event-related potential study. *Language, Cognition and Neuroscience*, 84–98.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314.

Fernando, T., & Kamp, H. (1996). Expecting Many. *Semantics and Linguistic Theory*, *6*, 53.

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry, N. W. (1983). Brain Potentials Related to Stages of Sentence Verification. *Psychophysiology*, *20*(4), 400–409.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge: MIT Press.

Glöckner, I. (2006). *Fuzzy Quantifiers: A Computational Theory*. Berlin, Heidelberg: Springer.

Greenberg, J. H. (1978). Generalizations about numeral systems. In J. H. Greenberg (Ed.), *Universals of Human Language* (pp. 250–295). Stanford: Stanford University Press.

Grice, P. H. (1975). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics* (pp. 41–58). New York: Academic Press.

Grisoni, L., Mccormick Miller, T., & Pulvermüller, F. (2017). Neural Correlates of Semantic Prediction and Resolution in Sentence Processing. *Journal of Neuroscience*, *37*(18), 4848–4858.

Grisoni, L., Tomasello, R., & Pulvermüller, F. (2021). Correlated Brain Indexes of Semantic Prediction and Prediction Error: Brain Localization and Category Specificity. *Cerebral Cortex*, *31*(3), 1553–1568.

Grodzinsky, Y., Agmon, G., Snir, K., Deschamps, I., & Loewenstein, Y. (2018). processing cost of Downward Entailingness: the representation and verification of comparative constructions. *ZAS Papers in Linguistics*, *60*, 435–451.

Grodzinsky, Y., Jaichenco, V., Deschamps, I., Sánchez, M. E., Fuchs, M., Pieperhoff, P., ... Amunts, K. (2020). Negation and the Brain. In V. Dèprez & T. M. Espinal (Eds.), *The Oxford Handbook of Negation* (pp. 693–712). Oxford University Press.

Guest, O., & Martin, A. (2021). How computational modeling can force theory building in psychological science. *Perspectives on psychological science : a journal of the Association for Psychological Science*, *16*(4), 789–802.

Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin and Review*, *26*(3), 772–789.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: Most versus more than half. *Natural Language Semantics*, *17*(1), 63–98.

Hagoort, P., Brown, C., & Groothusen, J. (1993). Language and Cognitive Processes The syntactic positive shift (sps) as an erp measure of syntactic processing The Syntactic Positive Shift (SPS) as an ERP Measure of Syntactic Processing. *Language and Cognitive Processes*, *8*(4), 439–483.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science*, *304*(5669), 438–441.

Hammerton, M. (1976). How much is a large part? *Applied Ergonomics*, *7*(1), 10–12.

Hanlbleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park: Sage Publications.

Harold Jeffreys. (1961). *The Theory of Probability* (3rd ed.). New York: Oxford University Press.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100.

Hebbali, A. (2020). Tools for Building OLS Regression Models [Computer software manual]. (R package version 0.5.3)

Heekeren, H. R., Marrett, S., Bandettini, P. A., & Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature*, *431*(7010), 859–862.

Heim, S., McMillan, C. T., Clark, R., Golob, S., Min, N. E., Olm, C., ... Grossman, M. (2015). If so many are "few", how few are "many"? *Frontiers in Psychology*, *6*.

Heim, S., Peiseler, N., & Bekemeier, N. (2020). "Few" or "Many"? An Adaptation Level Theory Account for Flexibility in Quantifier Processing. *Frontiers in Psychology*, *11*, 382.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: a new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–90.

Hockett, C. F. (1963). The problem of universals in language. In J. H. Greenberg (Ed.), *Universals of Language* (pp. 1–29). Cambridge, MA: MIT Press.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, *14*(4), 382–401.

Horn, L. R. (2001). *A Natural History of Negation.* CSLI.

Hunter, T., & Lidz, J. (2013). Conservativity and learnability of determiners. *Journal of Semantics*, *30*(3), 315–334.

Hyman, L. M. (2008). Universals in phonology. *Linguistic Review*, *25*(1-2), 83–137.

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, *10*(3), 244–253.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149.

Kang, I., & Ratcliff, R. (2020). Modeling the interaction of numerosity and perceptual variables with the diffusion model. *Cognitive Psychology*, *120*, 101288.

Katsimpokis, D., Hawkins, G. E., & van Maanen, L. (2020). Not all Speed-Accuracy Trade-Off Manipulations Have the Same Psychological Effect. *Computational Brain and Behavior*, *3*(3), 252–268.

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, *38*(7), 1033–1050.

Kaup, B., Ludtke, J., & Zwaan, R. A. (2007). The experiential view of language comprehension: How is negation represented? In F. Schmalhofer & C. A. Perfetti (Eds.), *Higher Level Language Processes in the Brain: Inference and Comprehension Processes* (pp. 255–288). London: Lawrence Erlbaum Associates.

Keenan, E. L., & Paperno, D. (2012). *Handbook of Quantifiers in Natural Language* (Vol. 90). Springer Science and Business Media B.V.

Keenan, E. L., & Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy*, *9*(3), 253–326.

Kelly, S. P., & O'Connell, R. G. (2015). The neural processes underlying perceptual decision making in humans: recent progress and future directions. *Journal of physiology, Paris*, *109*(1-3), 27–37.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual Differences in Language Acquisition and Processing. *Trends in Cognitive Sciences*, *22*(2), 154–169.

Kiefer, M., Marzinzik, F., Weisbrod, M., Scherg, M., & Spitzer, M. (1998). The time course of brain activations during response inhibition: evidence from event-related potentials in a go/no go task. *Neuroreport*, *9*(4), 765–770.

Kochari, A. R. (2019). Conducting web-based experiments for numerical cognition research. *Journal of Cognition*, *2*(1), 39.

Kolvoort, I., Davis, Z. J., van Maanen, L., & Rehder, B. (2021). Variability in Causal Judgments. In *Proceedings of 43rd the Annual Conference of the Cognitive Science Society* (pp. 1250–1256). Cognitive Science Society.

Kotek, H., Sudo, Y., & Hackl, M. (2015). Experimental investigations of ambiguity: the case of most. *Natural Language Semantics*, *23*(2), 119–156.

Kounios, J., & Holcomb, P. J. (1992). Structure and Process in Semantic Memory: Evidence From Event-Related Brain Potentials and Reaction Times. *Journal of Experimental Psychology: General*, *121*(4), 459–479.

Kusak, G., Grune, K., Hagendorf, H., & Metz, A. M. (2000). Updating of working memory in a running memory task: an event-related potential study. *International Journal of Psychophysiology*, *39*(1), 51–65.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13).

Leonhard, T., Fernández, S. R., Ulrich, R., & Miller, J. (2011). Dual-Task Processing When Task 1 Is Hard and Task 2 Is Easy: Reversed Central Processing Order? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 115–136.

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods*, *49*(2), 513–537.

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of most. *Natural Language Semantics*, *19*(3), 227–256.

Ligia, S., Bortolotti, V., Tezza, R., Francisco De Andrade, D., Bornia, A. C., Cezar Bornia, A., ... De Sousa Júnior, A. F. (2013). Relevance and advantages of using the item response theory. *Quality & Quantity*, *47*, 2341–2360.

Lindström, P. (1966). First Order Predicate Logic with Generalized Quantifiers. *Theoria*, *32*(3), 186–195.

Liu, T., & Pleskac, T. J. (2011). Neural correlates of evidence accumulation in a perceptual decision task. *Journal of Neurophysiology*, *106*(5), 2383–2398.

Liu, Y., & Kerre, E. E. (1998). An overview of fuzzy quantifiers. (I). Interpretations. *Fuzzy Sets and Systems*, *95*(1), 1-21.

Luck, S. J., Woodman, G. F., & Vogel, E. K. (2000). Event-related potential studies of attention. *Trends in Cognitive Sciences*, *4*(11), 432–440.

van Maanen, L., Portoles, O., & Borst, J. P. (2021). The Discovery and Interpretation of Evidence Accumulation Stages. *Computational Brain & Behavior 2021*, 1–21.

van Maanen, L., & van Rijn, H. (2010). The Locus of the Gratton Effect in Picture–Word Interference. *Topics in Cognitive Science*, *2*(1), 168–180.

van Maanen, L., van Rijn, H., & Taatgen, N. (2012). RACE/A: An Architectural Account of the Interactions Between Learning, Task Control, and Retrieval Dynamics. *Cognitive Science*, *36*(1), 62–101.

Makeig, S., Westerfield, M., Jung, T.-P., Enghoff, S., Townsend, J., Courchesne, E., & Sejnowski, T. (2002). Dynamic brain sources of visual evoked responses. (Reports). *Science*, *295*(5555), 690–695.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

Martin, A. E. (2016). Language Processing as Cue Integration: Grounding the Psychology of Language in Perception and Neurophysiology. *Frontiers in Psychology*, *7*, 120.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472.

McMillan, C. T., Clark, R., Moore, P., Devita, C., & Grossman, M. (2005). Neural basis for generalized quantifier comprehension. *Neuropsychologia*, *43*(12), 1729–1737.

Mhasawade, V., Szabó, I. E., Tosik, M., & Wang, S.-F. (2018). Neural Networks and Quantifier Conservativity: Does Data Distribution Affect Learnability? [Unpublished manuscript].

Miletić, S., & van Maanen, L. (2019). Caution in decision-making under time pressure is mediated by timing ability. *Cognitive Psychology*, *110*, 16–29.

Montague, R. (1970). Universal grammar. *Theoria*, *36*(3), 373–398.

Morey, R. (2018). *Package 'BayesFactor': Computation of Bayes Factors for Common Designs.* (R package version 0.9.12-4.3)

Mostowski, A. (1957). On a generalization of quantifiers. *Fundamenta Mathematicae*, *44*(1), 12–36.

Mulder, M. J., van Maanen, L., & Forstmann, B. U. (2014). Perceptual decision neurosciences - a model-based review. *Neuroscience*, *277*, 872–884.

Newstead, S. E., & Coventry, K. R. (2000). The role of context and functionality in the interpretation of quantifiers. *European Journal of Cognitive Psychology*, *12*(2), 243–259.

Newstead, S. E., Pollard, P., & Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Applied Ergonomics*, *18*(3), 178–182.

Nieuwland, M. S. (2016). Quantification, prediction, and the online impact of sentence truth-value: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning Memory and Cognition*, *42*(2), 316–334.

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, *63*(3), 324–346.

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213–1218.

Nordmeyer, A. E., & Frank, M. C. (2014). A pragmatic account of the processing of negative sentences. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society.* Austin TX: Cognitive Science Society.

Oaksford, M., Roberts, L., & Chater, N. (2002). Relative informativeness of quantifiers used in syllogistic reasoning. *Memory and Cognition*, *30*(1), 138–149.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*.

Orenes, I., Moxey, L., Scheepers, C., & Santamaría, C. (2016). Negation in context: Evidence from the visual world paradigm. *Quarterly Journal of Experimental Psychology*, *69*(6), 1082–1092.

Osth, A. F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology*, *92*, 101–126.

Palaz, B., Rhodes, R., & Hestvik, A. (2020). Informative use of "not" is N400-blind. *Psychophysiology*, *57*(12).

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 376–404.

Park, J., & Starns, J. J. (2015). The approximate number system acuity redefined: A diffusion model approach. *Frontiers in Psychology*, *6*, 1955.

Partee, B. (1989). Many Quantifiers. In J. Powers & K. de Jong (Eds.), *Escol 89: Proceedings of the Eastern States Conference on Linguistics* (pp. 241–258). Columbus, OH: Department of Linguistics, Ohio State University.

Pauli, P., Lutzenberger, W., Birbaumer, N., Rickard, T. C., & Bourne, L. E. (1996). Neurophysiological correlates of mental arithmetic. *Psychophysiology*, *33*(5), 522–529.

Peters, S., & Westerståhl, D. (2008). *Quantifiers in Language and Logic*. New York: Oxford University Press.

Petrov, A. A., van Horn, N. M., & Ratcliff, R. (2011). Dissociable perceptual-learning mechanisms revealed by diffusion-model analysis. *Psychonomic Bulletin and Review*, *18*(3), 490–497.

Pexman, P. M., & Yap, M. J. (2018). Individual differences in semantic processing: Insights from the Calgary Semantic Decision Project. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(7), 1091–1112.

Pezzelle, S., Bernardi, R., & Piazza, M. (2018). Probing the mental representation of quantifiers. *Cognition*, *181*, 117–126.

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of 'Most': Semantics, numerosity and psychology. *Mind and Language*, *24*(5), 554–585.

Pietroski, P., Lidz, J., Hunter, T., Odic, D., & Halberda, J. (2011). Seeing what you mean, mostly. *Syntax and Semantics*, *37*, 181–217.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421–425.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, *118*(10), 2128–2148.

van de Pol, I., Lodder, P., Maanen, L. v., Steinert-Threlkeld, S., & Szymanik, J. (2021). Quantifiers satisfying semantic universals are simpler. In *Proceedings of the 43rd Annual Conference of the Cognitive Science Society* (pp. 756–762). Cognitive Science Society.

van de Pol, I., Steinert-Threlkeld, S., & Szymanik, J. (2019). Complexity and learnability in the explanation of semantic universals of quantifiers. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society.* Montreal, QB: Cognitive Science Society.

Ramotowska, S., Archambeau, K., Augurzky, P., Schlotterbeck, F., Berberyan, H., van Maanen, L., & Szymanik, J. (2022). Discovering stages of processing in quantified sentences [Unpublished manuscript].

Ramotowska, S., Haaf, J., van Maanen, L., & Szymanik, J. (2022). Most quantifiers have many meanings [Unpublished manuscript].

Ramotowska, S., Steinert-Threlkeld, S., van Maanen, L., & Szymanik, J. (2020a). Individual differences in semantic representations: The case of most and more than half [Unpublished manuscript].

Ramotowska, S., Steinert-Threlkeld, S., van Maanen, L., & Szymanik, J. (2020b). Most, but not more than half, is proportion-dependent and sensitive to individual differences. In M. Franke, N. Kompa, M. Liu, J. L. Mueller, & J. Schwab (Eds.), *Proceedings of Sinn und Bedeutung 24* (Vol. 2, pp. 165–182). Osnabrück University and Humboldt University of Berlin.

Ramotowska, S., Steinert-Threlkeld, S., van Maanen, L., & Szymanik, J. (2021). Uncovering the structure of semantic representations using a computational model of decision-making [Unpublished manuscript].

Ramotowska, S., van Maanen, L., & Szymanik, J. (2022). Does ease of learning explain quantifier universals? [Unpublished manuscript].

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*(3), 446–461.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Ratcliff, R., & McKoon, G. (2018). Modeling numerosity representation with an integrated diffusion model. *Psychological Review*, *125*(2), 183–217.

Ratcliff, R., & McKoon, G. (2020). Decision making in numeracy tasks with spatially continuous scales. *Cognitive Psychology*, *116*, 101259.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*(3), 127–157.

Ratcliff, R., Thompson, C. A., & McKoon, G. (2015). Modeling individual differences in response time and accuracy in numeracy. *Cognition*, *137*, 115–136.

Regel, S., Meyer, L., & Gunter, T. C. (2014). Distinguishing neurocognitive processes reflected by P600 effects: Evidence from ERPs and neural oscillations. *PLoS ONE*, *9*(5), e96840.

Register, J., Mollica, F., & Piantadosi, S. T. (2018). Semantic verification is flexible and sensitive to context [Unpublished manuscript].

Roever, C., Raabe, N., Luebke, K., Ligges, U., Szepannek, G., & Zentgraf, M. (2015). *Package 'klaR': Classification and visualization.* (R package version 0.6-15)

Romoli, J. (2015). A Structural Account of Conservativity. *Semantics-Syntax Interface*, *2*, 28–57.

Ruchkin, D. S., Johnson, R., Grafman, J., Canoune, H., & Ritter, W. (1992). Distinctions and similarities among working memory processes: an event-related potential study. *Cognitive Brain Research*, *1*(1), 53–66.

Scaltritti, M., Job, R., Alario, F. X., & Sulpizio, S. (2020). On the boundaries between decision and action: Effector-selective lateralization of beta-frequency power is modulated by the lexical frequency of printed words. *Journal of Cognitive Neuroscience*, *32*(11), 2131–2144.

Schlotterbeck, F. (2017). *From truth conditions to processes: how to model the processing difficulty of quantified sentences based on semantic theory.*

Schlotterbeck, F., Ramotowska, S., van Maanen, L., & Szymanik, J. (2020). Representational complexity and pragmatics cause the monotonicity effect. In S. Denison, M. Mack, Y. Xu, & B. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3398–3404). Cognitive Science Society.

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, *43*(4), 441–464.

Schöller, A., & Franke, M. (2016). How many manys? Exploring semantic theories with data-driven computational models. In N. Bade, P. Berezovskaya, & A. Schöller (Eds.), *Proceedings of Sinn und Bedeutung 20* (pp. 622–639).

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461 - 464.

Shah, A. S., Bressler, S. L., Knuth, K. H., Ding, M., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2004). Neural Dynamics and the Fundamental Mechanisms of Event-related Brain Potentials. *Cerebral Cortex*, *14*(5), 476–483.

Shikhare, S., Heim, S., Klein, E., Huber, S., & Willmes, K. (2015). Processing of Numerical and Proportional Quantifiers. *Cognitive Science*, *39*(7), 1504–1536.

Solt, S. (2011). Vagueness in Quantity: Two Case Studies from a Linguistic Perspective. *Understanding vagueness. logical, philosophical and linguistic perspectives*, *36*, 157–174.

Solt, S. (2015). Vagueness and Imprecision: Empirical Foundations. *Annual Review of Linguistics*, *1*(1), 107–127.

Solt, S. (2016). On measurement and quantification: The case of most and more than half. *Language*, *92*(1), 65–100.

Spenader, J., & de Villiers, J. (2019). Are conservative quantifiers easier to learn? Evidence from novel quantifier experiments. In J. J. Schlöder, D. McHugh, & F. Roelofsen (Eds.), *Proceedings of the 22nd Amsterdam Colloquium* (pp. 504–512). University of Amsterdam.

Spychalska, M., Kontinen, J., & Werning, M. (2016). Investigating scalar implicatures in a truth-value judgement task: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, *31*(6), 817–840.

Stan Development Team. (2017). Shinystan: interactive visual and numerical diagnostics and posterior analysis for Bayesian models [Computer software manual]. (R package version 2.5.0)

Stan Development Team. (2020). RStan: the R interface to Stan [Computer software manual]. (R package version 2.21.3)

Steinert-Threlkeld, S. (2020). Quantifiers in natural language optimize the simplicity/informativeness trade-off. In J. J. Schlöder, D. McHugh, & F. Roelofsen (Eds.), *Proceedings of the 22nd Amsterdam Colloquium* (pp. 513–522). University of Amsterdam.

Steinert-Threlkeld, S., Munneke, G.-J., & Szymanik, J. (2015). Alternative Representations in Formal Semantics: A case study of quantifiers. In T. Brochhagen, F. Roelofsen, & N. Theiler (Eds.), *Proceedings of the 20th Amsterdam Colloquium* (pp. 368–378). University of Amsterdam.

Steinert-Threlkeld, S., & Szymanik, J. (2019). Learnability and semantic universals. *Semantics and Pragmatics*, *12*(4), 1.

Steinert-Threlkeld, S., & Szymanik, J. (2020). Ease of learning explains semantic universals. *Cognition*, *195*, 104076.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*(C), 276–315.

Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2011). Language-Related ERP Components. In E. S. Kappenman & S. J. Luck (Eds.), *The Oxford Handbook of Event-Related Potential Components.* Oxford University Press.

Szabolcsi, A. (2010). *Quantification.* Cambridge University Press.

Szymanik, J. (2016). *Quantifiers and Cognition. Logical and Computational Perspectives.* Springer.

Szymanik, J., & Zajenkowski, M. (2010). Comprehension of simple quantifiers: Empirical evaluation of a computational model. *Cognitive Science*, *34*(3), 521–532.

Szymanik, J., & Zajenkowski, M. (2013). Monotonicity has only a relative effect on the complexity of quantifier verification. In M. Aloni, M. Franke, & F. Roelofsen (Eds.), *Proceedings of the 19th Amsterdam Colloquium* (pp. 219–225). University of Amsterdam.

Talmina, N., Kochari, A., & Szymanik, J. (2017). Quantifiers and verification strategies: connecting the dots. In A. Cremers, T. van Gessel, & F. Roelofsen (Eds.), *Proceedings of the 21st Amsterdam Colloquium* (pp. 465–473). University of Amsterdam.

Thornhill, D. E., & van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, *83*(3), 382–392.

Tian, Y., Breheny, R., & Ferguson, H. J. (2010). Why we simulate negated information: A dynamic pragmatic account. *Quarterly Journal of Experimental Psychology*, *63*(12), 2305–2312.

van Tiel, B., Franke, M., & Sauerland, U. (2021). Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(9).

Tillman, G., Osth, A. F., van Ravenzwaaij, D., & Heathcote, A. (2017). A diffusion decision model analysis of evidence variability in the lexical decision task. *Psychonomic Bulletin & Review*, *24*(6), 1949–1956.

Tucker, D., Tomaszewicz, B., & Wellwood, A. (2018). Decomposition and Processing of Negative Adjectival Comparatives. In E. Castroviejo, L. McNally, & G. Weidman Sassoon (Eds.), *The Semantics of Gradability, Vagueness, and Scale Structure: Experimental Perspectives* (Vol. 4, pp. 243–273). Springer.

Twomey, D. M., Murphy, P. R., Kelly, S. P., & O'Connell, R. G. (2015). The classic P300 encodes a build-to-threshold decision variable. *The European Journal of Neuroscience*, *42*(1), 1636–1643.

Urbach, T. P., DeLong, K. A., & Kutas, M. (2015). Quantifiers are incrementally interpreted in context, more than less. *Journal of Memory and Language*, *83*, 79–96.

Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, *63*(2), 158–179.

Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). Computational Models of Retrieval Processes in Sentence Processing. *Trends in Cognitive Sciences*, *23*(11), 968–982.

Verheyen, S., Dewil, S., & Égré, P. (2018). Subjectivity in gradable adjectives: The case of tall and heavy. *Mind & Language*, *33*(5), 460–479.

Verheyen, S., Droeshout, E., & Storms, G. (2019). Age-Related Degree and Criteria Differences in Semantic Categorization. *Journal of Cognition*, *2*(1).

Verheyen, S., & Storms, G. (2013). A Mixture Approach to Vagueness and Ambiguity. *PLoS ONE*, *8*(5), e63507.

Verheyen, S., & Storms, G. (2018). Education as a Source of Vagueness in Criteria and Degree. In E. Castroviejo, L. McNally, & G. Sassoon (Eds.), *The Semantics of Gradability, Scale Structure and Vagueness: Experimental Perspectives* (pp. 149–167). Springer.

Verheyen, S., White, A., & Égré, P. (2019). Revealing Criterial Vagueness in Inconsistencies. *Open Mind*, *3*, 41–51.

Wagenmakers, E. J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*(5), 641–671.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin and Review*, *11*(1), 192–196.

Waldon, B., & Degen, J. (2020). Modeling Behavior in Truth Value Judgment Task Experiments. In A. Ettinger, G. Jarosz, & M. Nelson (Eds.), *Proceedings of the Society for Computation in Linguistics* (Vol. 3, pp. 238–247). Society for Computation in Linguistics.

Wason, P. C. (1961). Response to Affirmative and Negative Binary Statements. *British Journal of Psychology*, *52*(2), 133–142.

Wiswede, D., Koranyi, N., Müller, F., Langner, O., & Rothermund, K. (2013). Validating the truth of propositions: Behavioral and ERP indicators of truth evaluation processes. *Social Cognitive and Affective Neuroscience*, *8*(6), 647–653.

Xiang, M., Kramer, A., & Nordmeyer, A. E. (2020). An informativity-based account of negation complexity. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*(10), 1857–1867.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53–79.

Yeung, N., Bogacz, R., Holroyd, C. B., & Cohen, J. D. (2004). Detection of synchronized oscillations in the electroencephalogram: An evaluation of methods. *Psychophysiology*, *41*(6), 822–832.

Yeung, N., Bogacz, R., Holroyd, C. B., Nieuwenhuis, S., & Cohen, J. D. (2007). Theta phase resetting and the error-related negativity. *Psychophysiology*, *44*(1), 39–49.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, *87*, 128–143.

Young, R., & Chase, W. G. (1971). Additive stages in the comparison of sentences and pictures. In *Midwestern Psychological Association Meetings.* Chicago.

Zadeh, L. A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications*, *9*(1), 149–184.

Zajenkowski, M., & Szymanik, J. (2013). MOST intelligent people are accurate and SOME fast people are intelligent. Intelligence, working memory, and semantic processing of quantifiers from a computational perspective. *Intelligence. A Multidisciplinary Journal*, *41*(5), 456–466.

Zajenkowski, M., Szymanik, J., & Garraffa, M. (2014). Working Memory Mechanism in Proportional Quantifier Verification. *Journal of Psycholinguistic Research*, *43*, 839–853.

Zhang, Q., van Vugt, M., Borst, J. P., & Anderson, J. R. (2018). Mapping working memory retrieval in space and in time: A combined electroencephalography and electrocorticography approach. *NeuroImage*, *174*, 472–484.

Zhang, Q., Walsh, M. M., & Anderson, J. R. (2017). The effects of probe similarity on retrieval and comparison processes in associative recognition. *Journal of Cognitive Neuroscience*, *29*(2), 352–367.

Zhang, Q., Walsh, M. M., & Anderson, J. R. (2018). The Impact of Inserting an Additional Mental Process. *Computational Brain & Behavior*, *1*(1), 22–35.

Zhao, M., Liu, T., Chen, G., & Chen, F. (2015). Are scalar implicatures automatically processed and different for each individual? A mismatch negativity (MMN) study. *Brain Research*, *1599*, 137–49.

Zylberberg, A., Dehaene, S., Roelfsema, P. R., & Sigman, M. (2011). The human Turing machine: a neural framework for mental programs. *Trends in Cognitive Sciences*, *15*(7), 293–300.

Samenvatting

Het kwantificeren van de representatie van kwantoren: Experimentele studies, computationele modellen en individuele verschillen

Dit proefschrift stelt een nieuw, cognitief perspectief voor op de representatie en verificatie van betekenis van kwantoren in natuurlijke taal. Volgens het traditionele, logische perspectief worden dergelijke uitdrukkingen gerepresenteerd door middel van de voorwaarden waaronder de uitdrukking waar is. De aanname is dat deze waarheidsvoorwaarden gelden voor alle gebruikers van een taal. Echter, meer en meer onderzoek laat zien dat er wel degelijk variabiliteit is tussen representatie en verificatie van betekenis van kwantorenvan individuen. Dergelijke indivduele verschillen kunnen niet verklaard worden door het logische perspectief.

Het cognitieve perspectief laat dergelijke variatie tussen individuen wel toe. Daarnaast biedt dit persepctief een verklaring voor formele eigenschappen van kwantoren, zoals vaagheid en polariteit

Om de representatie van betekenis te onderzoeken, spelen computationele modellen een belangrijke rol. In dit proefschirft worden drie verschillende computationele modellen besproken. Elk model vat verschillende aspecten van representatie en verificatie van de betekenis van gekwantificeerde zinnen, zoals de waarheidsvoorwaarden, de vaagheid, of de cognitieve processen. Daarnaast beiden de computationele modellen de mogelijkheid om formeleeigenschappen van kwantoren te onderscheiden van individuele verschillen in taakvaardigheid.

Hoofdstuk 2 van dit proefschrift onderzoekt individuele verschilllen in betekenisrepresentatie van vijf kwantoren die in het Engels voorkomen: *few*, *many*, *most*, *fewer than half* en *more than half*. Door middel van een computationeel model vatten we twee eigenschappen van betekenis van kwantoren de waarheidsvoorwaarde, de vaagheid, en onderscheiden we dit van variatie in taakvaardigheid van proefpersonen (responsie fouten). Door middel van een clusteranalyse laten we zien dat er drie groepen proefpersonen zijn met verschillende waarheidsvoorwaarde voor textitfew, *many* en *most*. Deze groepen verschillen ook met be-

trekking tot de vaagheid die ze aan de kwantoren toekennen en de ordening van kwantoren de mentale getallenlijn.

Hoofdstuk 3 verbreedt de resultaten van Hoofdstuk 2. Door gebruik te maken van een complexer computational model focussen we hier op de kwantoren *most* en *more than half*. De resultaten laten zien dat *most* gevoelig is voor individuele verschillen in betekenisrepresentatie, en dat de verificatie van zinnen met *most* afhangt van proportie. Een extra resultaat van Hoodstuk 3 is de observatie dat ondanks de betekenisverschillen tussen individueen, de betekenisrepresentatie wel stabiel op verschillende tijdsmomenten.

Positieve kwantoren (zoals *more than half*) worden sneller verwerkt dan hun negatieve tegenhangers (zoals *fewer than half*). In Hoofdstuk 4 testen we de voorspellingen van twee verschillende theorieen van dit zogenaamde *polariteitseffect*. Dit zijn de pragmatische theorie, en het "tweetrapsmodel". Twee experimenten en computationele modellen laten zien dat het polariteitseffecten twee verschillende oorzaken kent. Beide theorieen lijken dus ondersteund te worden.

In Hoofdstuk 5 gaan we dieper in op de oorzaken van het polariteitseffect door een belangrijke voorspelling van het tweetrapsmodel direct te toetsen. Het tweetrapsmodel stelt dat zinnen met negatieve kwantoren langzamer geverifieerd worden omdat ze een extra verwerkingsstap nodig hebben ten opzichte van positieve kwantoren. Hoofdstuk 5 presenteert de resultaten van een elektro-encefalografiestudie waarim proefpersonen de waarheid van zinnen moesten verifieren aan de hand van een afbeelding. De zinnen hadden of een postieve of een negatieve kwantor. Door het aantal verwerkingsstappen te schatten op basis van het EEG patroon, vonden we geen extra verwerkingsstap voor negatieve kwantoren, wat niet consistent is met het tweetrapsmodel.

Hoofdstuk 6 tenslotte onderzoekt een verklaring voor het bestaan van semantische universalia, die belangrijk zijn voor de acquisitie begrip van kwantoren. Het hoofdstuk focust op de hypothese dat nieuwe kwantoren die voldoen aan de semantische unversalia makkelijker te leren zijn. In een groot experiment testen we de leercurve van acht verschillende kwantoren, die verschillen wat betreft *monotoniciteit*, *conservatisme* en *kwantiteit*. De resultaten ondersteunen de hypothese dat althans sommige sematnische unversalia makkelijker te leren zijn. Daarnaast bevat dit hoofdstuk ook een belangrijke discussie van methodologische overwegingen bji het experimentele onderzoek van semantische unversalia.

Abstract

Quantifying quantifier representations: Experimental studies, computational modeling, and individual differences

This thesis proposes a new, cognitive perspective on the meaning representations and verification of natural language quantifiers. According to the traditional, logical view, quantity words are represented in the form of truth conditions shared across language users. However, a growing body of evidence shows variability among speakers in semantic representations and verification strategies of quantifies. The logical view cannot explain the individual differences in meaning representations of quantity words.

In contrast, according to the cognitive perspective, the truth-conditional representations of quantifiers may vary between speakers. Moreover, the model captures the properties of quantifiers such as vagueness and polarity.

Computational models play a key role in the investigation of meaning representations. This thesis presents three computational models. Each model captures different aspects of the representation and verification of quantified sentences, for example, the quantifier's truth condition, vagueness, or processing stages. Moreover, computational models disentangle the formal properties of quantifiers from individual differences in task performance.

Chapter 2 investigates the individual differences in meaning representations of five natural language quantifiers: *few*, *many*, *most*, *fewer than half*, and *more than half*. By using the computational model, we capture two key properties of quantifier meaning – truth conditions, vagueness – as well as variation in the task performance (response errors) of participants. The results of cluster analysis show that participants constitute three groups with different thresholds for *few*, *many*, and *most*. Moreover, the groups differ in the perception of the vagueness of quantifiers, and they put quantifiers in a different order on a mental line.

Chapter 3 further extends the findings of Chapter 2. By using another computational model, we investigate the meaning representations and verification of

most and *more than half*. The results of Chapter 3 show that *most* is sensitive to individual differences in representations and its verification is proportion-dependent. In addition, despite the individual differences in meanings, the meaning representations are stable over time.

Positive quantifiers (e.g., *more than half*) are processed faster than their negative counterparts. In Chapter 4, we use computational model to test the predictions of two competing accounts (pragmatic and two-step models) explaining the polarity effect. Two quantifier verification experiments and modeling data show two separate sources of polarity effect. In conclusion, the findings support both pragmatic and two-step accounts.

Chapter 5 further investigates the source of the polarity effect by directly testing the predictions of the two-step model. The two-step model postulates that the negative quantifiers are verified slower because they require and extra processing step compared to positive quantifiers. Chapter 5 presents the results of the electroencephalography picture-sentence verification experiment with two quantifiers: *fewer than half* and *more than half*. We used computational model to estimate and compare the number of processing stages of the quantified sentences. The findings of Chapter 5 challenge the two-step model.

Chapter 6 investigates one of the explanations of the semantic universals, namely the learnability hypothesis. In a large-scale experiment, we test the speed of acquisition of eight different quantifiers that vary in three formal properties: monotonicity, conservativity, and quantity. The findings of Chapter 6 support the learnability explanation of some of the semantic universals. Moreover, Chapter 6 stresses methodological aspects of experimental investigation of semantic universals.

Titles in the ILLC Dissertation Series:

ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion

ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm

ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models

ILLC DS-2016-05: **Phong Lê**
Learning Vector Representations for Sentences - The Recursive Deep Learning Approach

ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
Aligning the Foundations of Hierarchical Statistical Machine Translation

ILLC DS-2016-07: **Andreas van Cranenburgh**
Rich Statistical Parsing and Literary Language

ILLC DS-2016-08: **Florian Speelman**
Position-based Quantum Cryptography and Catalytic Computation

ILLC DS-2016-09: **Teresa Piovesan**
Quantum entanglement: insights via graph parameters and conic optimization

ILLC DS-2016-10: **Paula Henk**
Nonstandard Provability for Peano Arithmetic. A Modal Perspective

ILLC DS-2017-01: **Paolo Galeazzi**
Play Without Regret

ILLC DS-2017-02: **Riccardo Pinosio**
The Logic of Kant's Temporal Continuum

ILLC DS-2017-03: **Matthijs Westera**
Exhaustivity and intonation: a unified theory

ILLC DS-2017-04: **Giovanni Cinà**
Categories for the working modal logician

ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
Communication and Computation: New Questions About Compositionality

ILLC DS-2017-06: **Peter Hawke**
The Problem of Epistemic Relevance

ILLC DS-2017-07: **Aybüke Özgün**
Evidence in Epistemic Logic: A Topological Perspective

ILLC DS-2017-08: **Raquel Garrido Alhama**
Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ILLC DS-2017-09: **Miloš Stanojević**
Permutation Forests for Modeling Word Order in Machine Translation

ILLC DS-2018-01: **Berit Janssen**
Retained or Lost in Transmission? Analyzing and Predicting Stability in Dutch Folk Songs

ILLC DS-2018-02: **Hugo Huurdeman**
Supporting the Complex Dynamics of the Information Seeking Process

ILLC DS-2018-03: **Corina Koolen**
Reading beyond the female: The relationship between perception of author gender and literary quality

ILLC DS-2018-04: **Jelle Bruineberg**
Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems

ILLC DS-2018-05: **Joachim Daiber**
Typologically Robust Statistical Machine Translation: Understanding and Exploiting Differences and Similarities Between Languages in Machine Translation

ILLC DS-2018-06: **Thomas Brochhagen**
Signaling under Uncertainty

ILLC DS-2018-07: **Julian Schlöder**
Assertion and Rejection

ILLC DS-2018-08: **Srinivasan Arunachalam**
Quantum Algorithms and Learning Theory

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**
Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks

ILLC DS-2018-10: **Chenwei Shi**
Reason to Believe

ILLC DS-2018-11: **Malvin Gattinger**
New Directions in Model Checking Dynamic Epistemic Logic

ILLC DS-2018-12: **Julia Ilin**
Filtration Revisited: Lattices of Stable Non-Classical Logics

ILLC DS-2018-13: **Jeroen Zuiddam**
Algebraic complexity, asymptotic spectra and entanglement polytopes

ILLC DS-2019-01: **Carlos Vaquero**
What Makes A Performer Unique? Idiosyncrasies and commonalities in expressive music performance

ILLC DS-2019-02: **Jort Bergfeld**
Quantum logics for expressing and proving the correctness of quantum programs

ILLC DS-2019-03: **András Gilyén**
Quantum Singular Value Transformation & Its Algorithmic Applications

ILLC DS-2019-04: **Lorenzo Galeotti**
The theory of the generalised real numbers and other topics in logic

ILLC DS-2019-05: **Nadine Theiler**
Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles

ILLC DS-2019-06: **Peter T.S. van der Gulik**
Considerations in Evolutionary Biochemistry

ILLC DS-2019-07: **Frederik Möllerström Lauridsen**
Cuts and Completions: Algebraic aspects of structural proof theory

ILLC DS-2020-01: **Mostafa Dehghani**
Learning with Imperfect Supervision for Language Understanding

ILLC DS-2020-02: **Koen Groenland**
Quantum protocols for few-qubit devices

ILLC DS-2020-03: **Jouke Witteveen**
Parameterized Analysis of Complexity

ILLC DS-2020-04: **Joran van Apeldoorn**
A Quantum View on Convex Optimization

ILLC DS-2020-05: **Tom Bannink**
Quantum and stochastic processes

ILLC DS-2020-06: **Dieuwke Hupkes**
Hierarchy and interpretability in neural models of language processing

ILLC DS-2020-07: **Ana Lucia Vargas Sandoval**
On the Path to the Truth: Logical & Computational Aspects of Learning

ILLC DS-2020-08: **Philip Schulz**
Latent Variable Models for Machine Translation and How to Learn Them

ILLC DS-2020-09: **Jasmijn Bastings**
A Tale of Two Sequences: Interpretable and Linguistically-Informed Deep Learning for Natural Language Processing

ILLC DS-2020-10: **Arnold Kochari**
Perceiving and communicating magnitudes: Behavioral and electrophysiological studies

ILLC DS-2020-11: **Marco Del Tredici**
Linguistic Variation in Online Communities: A Computational Perspective

ILLC DS-2020-12: **Bastiaan van der Weij**
Experienced listeners: Modeling the influence of long-term musical exposure on rhythm perception

ILLC DS-2020-13: **Thom van Gessel**
Questions in Context

ILLC DS-2020-14: **Gianluca Grilletti**
Questions & Quantification: A study of first order inquisitive logic

ILLC DS-2020-15: **Tom Schoonen**
Tales of Similarity and Imagination. A modest epistemology of possibility

ILLC DS-2020-16: **Ilaria Canavotto**
Where Responsibility Takes You: Logics of Agency, Counterfactuals and Norms

ILLC DS-2020-17: **Francesca Zaffora Blando**
Patterns and Probabilities: A Study in Algorithmic Randomness and Computable Learning

ILLC DS-2021-01: **Yfke Dulek**
Delegated and Distributed Quantum Computation

ILLC DS-2021-02: **Elbert J. Booij**
The Things Before Us: On What it Is to Be an Object

ILLC DS-2021-03: **Seyyed Hadi Hashemi**
Modeling Users Interacting with Smart Devices

ILLC DS-2021-04: **Sophie Arnoult**
Adjunction in Hierarchical Phrase-Based Translation

ILLC DS-2021-05: **Cian Guilfoyle Chartier**
A Pragmatic Defense of Logical Pluralism

ILLC DS-2021-06: **Zoi Terzopoulou**
Collective Decisions with Incomplete Individual Opinions

ILLC DS-2021-07: **Anthia Solaki**
Logical Models for Bounded Reasoners

ILLC DS-2021-08: **Michael Sejr Schlichtkrull**
Incorporating Structure into Neural Models for Language Processing

ILLC DS-2021-09: **Taichi Uemura**
Abstract and Concrete Type Theories

ILLC DS-2021-10: **Levin Hornischer**
Dynamical Systems via Domains: Toward a Unified Foundation of Symbolic and Non-symbolic Computation

ILLC DS-2021-11: **Sirin Botan**
Strategyproof Social Choice for Restricted Domains

ILLC DS-2021-12: **Michael Cohen**
Dynamic Introspection

ILLC DS-2022-01: **Anna Bellomo**
Sums, Numbers and Infinity: Collections in Bolzano's Mathematics and Philosophy

ILLC DS-2022-02: **Jan Czajkowski**
Post-Quantum Security of Hash Functions