From truth conditions to processes: how to model the processing difficulty of quantified sentences based on semantic theory

Dissertation zur

Erlangung des akademischen Grades

Doktor der Philosophie

in der Philosophischen Fakultät

der

vorgelegt von

Fabian Schlotterbeck

aus
Tübingen

2017

Gedruckt mit Genehmigung der Philosophischen Fakultät der
Eberhard Karls Universität Tübingen

Dekan:	Prof. Dr. Jürgen Leonhardt

Hauptberichterstatter:	Prof. Dr. Fritz Hamm
Mitberichterstatter:	Prof. Dr. Rolf Ulrich
	Prof. Dr. Wolfgang Sternefeld

Tag der mündlichen Prüfung: 14. Juli 2017

Universitätsbibliothek Tübingen
Online-Bibliotheksinformations- und Ausleihsystem TOBIAS-lib

ABSTRACT

The present dissertation is concerned with the processing difficulty of quantified sentences and how it can be modeled based on semantic theory. Processing difficulty of quantified sentences is assessed using psycholinguistic methods such as systematically collecting truth-value judgments or recording eye movements during reading. Predictions are derived from semantic theory via parsimonious processing assumptions, taking into account automata theory, signal detection theory and computational complexity.

Chapter 1 provides introductory discussion and overview. Chapter 2 introduces basic theoretical concepts that are used throughout the rest of the dissertation. In chapter 3, processing difficulty is approached on an abstract level. The difficulty of the truth evaluation of reciprocal sentences with generalized quantifiers as antecedents is classified using computational complexity theory. This is independent of the actual algorithms or procedures that are used to evaluate the sentences. One production and one sentence-picture verification experiment are reported which tested whether cognitive capacities are limited to those functions that are computationally tractable. The results indicate that intractable interpretations occur in language comprehension but also that their verification rapidly exceeds cognitive capacities in case the verification problem cannot be solved using simple heuristics. Chapter 4 discusses two common approaches to model the canonical verification procedures associated with quantificational sentences. The first is based on the semantic automata model which conceives of quantifiers as decision problems and characterizes the computational resources that are needed to solve them. The second approach is based on the interface transparency thesis, which stipulates a transparent interface between semantic representations and the realization of verification procedures in the general cognitive architecture. Both approaches are evaluated against experimental data. Chapter 5 focuses on a test case that is challenging for both of these approaches. In particular, increased processing difficulty of *more than n* as compared to *fewer than n* is investigated. A processing model is proposed which integrates insights from formal semantics with models from cognitive psychology. This model can be seen as im-

plementation and extension of the interface transparency thesis. The truth evaluation process is conceived of as a stochastic process as described in sequential sampling models of decision making. The increased difficulty of *fewer than n* as compared to *more than n* is attributed to an extra processing step of scale-reversal that precedes the actual decision process. Predictions of the integrated processing model are tested and confirmed in two sentence-picture verification experiments. Chapter 6 discusses whether and how the integrated processing model can be extended to other quantifiers. An extension to proportional comparative quantifiers, like *fewer than half* and *more than half* is proposed and discussed in the light of existing experimental data. Moreover, it is shown that what are called empty-set effects can be naturally derived from the model. Chapter 6 presents data from two eye tracking experiments that show that *fewer than* leads to increased difficulty as compared to *more than* already during reading. Moreover, this effect is magnified if such quantifiers are combined with overt negation. Potential accounts of these findings are discussed. Conclusions are summarized in chapter 8.

ZUSAMMENFASSUNG

Die vorliegende Dissertation beschäftigt sich mit der Verarbeitungsschwierigkeit quantifizierter Sätze und damit, wie diese ausgehend von semantischer Theorie modelliert werden kann. Die Verarbeitungsschwierigkeit quantifizierter Sätze wird mittels psycholinguistischer Methoden, wie der systematischen Erhebung von Wahrheitswerturteilen oder dem Aufzeichnen von Blickbewegungen während des Lesens, ermittelt. Vorhersagen werden aus der semantischen Theorie mittels sparsamer Verarbeitungsannahmen abgeleitet. Dabei werden die Automatentheorie, die Signalentdeckungstheorie und die Komplexitätstheorie berücksichtigt.

Kapitel 1 bietet eine einleitende Diskussion und einen Überblick. Kapitel 2 führt grundlegende theoretische Konzepte ein, die über den Rest der Dissertation hinweg verwendet werden. In Kapitel 3 wird Verarbeitungsschwierigkeit auf einer abstrakten Ebene angegangen. Die Schwierigkeit der Wahrheitsbewertung von reziproken Sätzen mit generalisierten Quantoren als Antezedens wird mit Hilfe der Komplexitätstheorie klassifiziert. Dies geschieht unabhängig von den Algorithmen oder Prozeduren, die tatsächliche verwendet werden, um die Sätze zu evaluieren. Ein Produktions- und ein Satz-Bild-Verifikations-Experiment werden berichtet, welche untersuchten, ob kognitive Fähigkeiten auf solche Funktionen begrenzt sind, die effizient berechnet werden können. Die Resultate zeigen, dass Interpretationen, die nicht effizient berechenbar sind, zwar beim Sprachverstehen vorkommen aber auch, dass deren Verifikation rasch kognitive Fähigkeiten überschreitet, falls das Verifikationsproblem nicht durch simple Heuristiken gelöst werden kann. Kapitel 4 diskutiert zwei gängige Ansätze, die kanonischen Verifikationsprozeduren, die mit quantifizierten Sätzen verbunden sind, zu modellieren. Der erste basiert auf dem Modell der semantischen Automaten, welches Quantoren als Entscheidungsprobleme betrachtet und die Rechenressourcen charakterisiert, die zu deren Lösung notwendig sind. Der zweite Ansatz basiert auf der These der Schnittstellen-Transparenz, die eine durchsichtige Schnittstelle zwischen semantischen Repräsentationen und der Realisierung von Verifikationsprozeduren innerhalb der allgemeinen kognitiven Architektur stipuliert. Beide Ansätze werden

anhand experimenteller Befunde bewertet. Kapitel 5 konzentriert sich auf einen Beispielsfall der für beide diese Ansätze eine Herausforderung darstellt. Und zwar wird dort die erhöhte Verarbeitungsschwierigkeit von *mehr als n* im Vergleich zu *weniger als n* untersucht. Es wird ein Verarbeitungsmodell vorgeschlagen, welches Erkenntnisse aus der formalen Semantik mit Modellen aus der Kognitionspsychologie integriert. Dieses Modell kann als Implementierung und Erweiterung der These der Schnittstellen-Transparenz gesehen werden. Der Prozess der Wahrheitsbewertung wird als stochastischer Prozess verstanden, ähnlich sogenannter *sequential sampling* Modelle, also solcher, die Entscheidungsprozesse mittels sequentieller probabilistischer Methoden beschreiben. Die erhöhte Schwierigkeit von *weniger als n* im Vergleich zu *mehr als n* wird einem zusätzlichen Verarbeitungsschritt der Skalen-Umkehr zugeschrieben, welcher dem eigentlichen Entscheidungsprozess vorangeht. Vorhersagen dieses integrierten Modells werden in zwei Satz-Bild-Verifikations-Experimenten getestet und bestätigt. In Kapitel 6 wird diskutiert, ob und wie das vorgeschlagene integrierte Modell auf andere Quantoren erweitert werden kann. Eine Erweiterung auf proportionale komparative Quantoren, wie z.B. *weniger als die Hälfte* and *mehr als die Hälfte* wird vorgeschlagen und im Lichte existierender experimenteller Befunde diskutiert. Weiterhin wird gezeigt, dass sogenannte *empty-set effects* auf natürliche Weise aus dem vorgeschlagenen Modell abgeleitet werden können. Kapitel 6 beschreibt die Daten aus zwei Blickbewegungsverfolgungs-Experimenten, die zeigen, dass *weniger als* schon beim Lesen zu erhöhter Schwierigkeit im Vergleich zu *mehr als* führt. Zusätzlich wird dieser Effekt verstärkt, wenn solche Quantoren mit overter Negation kombiniert werden. Mögliche Erklärungen dieser Befunde werden diskutiert. Die Schlussfolgerungen werden in Kapitel 8 zusammengefasst.

DEDICATION

To Elin and Simon

ACKNOWLEDGEMENTS

Many people deserve my gratitude. Without their support, writing this dissertation would not have been possible. First and foremost, I would like to thank my advisors Fritz Hamm, Wolfgang Sternefeld and Rolf Ulrich and also Oliver Bott, who supported me from day one and were always there to answer my questions.

I would also like to thank my colleagues at the SFB833 and the University of Tübingen, especially Petra Augurzky, Tilman Berger, Sigrid Beck, Martin Butz, Verena Eikmeier, Michael Franke, Fritz Günther, Vera Hohaus, Robin Hörnig, Gerhard Jäger, Barbara Kaup, Andreas Konietzko, Hartmut Leuthold, Chris Miller, Andreas Nieder, Anna Pryslopska, Aysenur Sarcan, Anthea Schöller, Beate Starke, Arnim von Stechow, Britta Stolterfoht and everyone who participated in the psychlinguistics lunch or the *Doktoranden Kolloquium* at the SFB833.

Outside of Tübingen, my gratitude goes especially to Lucas Champollion, Berry Claus, Jakub Dotlačil, Lyn Frazier, Nina Gierasimczuk, Juhani Järvikivi, Udo Klein, Dan Lassiter, Yaron McNabb, Doris Penka, Iris van Rooij, Christoph Schepers, Torgrim Solstad, Stephanie Solt, Jakub Szymanik, Barbara Tomaszewicz, Dag Westerståhl and the participants of the 3rd Workshop on Semantic Processing, Logic and Cognition in Tübingen, 2011, the Logic and Cognition Workshop at ESSLLI, 2012 and the Experimental Approaches to Semantics Workshop at ESSLLI, 2015.

Last but not least, I am grateful for the love and support of my family and friends, above all Olga, Raoul, Anja, Ulf and of course Elin and Simon.

CONTENTS

LIST OF TABLES AND FIGURES

ABBREVIATIONS

SYNTACTIC CATEGORIES

Note. *Syntactic categories are generally used without definition. There may be some inconsistencies in use. The categories* NP *and* DP *are used interchangeably. The question which is to be used when is independent of the contents of this work.*

ADJ	adjective
ADJP	adjective phrase
AUX	auxiliary
CNP	common noun phrase
C	complementizer
CP	complementizer phrase
DET	determiner
DEG	degree
DEGP	degree phrase
DP	determiner phrase
IP	inflection phrase
I	inflection
NEG	negation
N	noun
NP	noun phrase
PRO	pronoun
P	preposition
PP	preposition phrase

Q	quantifier
S	sentence
V	verb
VP	verb phrase

acronyms

2AFC	two alternative forced choice	
ANS	approximate number system	(e.g. Section 4.2.2.1)
CE	CONS and EXT	
CONS	conservativity	(Definition 2:10)
CCG	combinatory categorial grammar	(Section 2.7)
DDM	drift diffusion model	(Section 2.5.2)
DE	downward entailing	(Definition 2:8)
DFA	deterministic finite state automaton	(Definition 2:15)
EEG	electroencephalography	
ERP	event related potential	
EXT	domain independent / extension	(Definition 2:9)
fMRI	functional magnetic resonance imaging	
FPT	fixed parameter tractable	
GQ	generalized quantifier	(Definition 2:1)
GQT	generalized quantifier theory	(Section 2.2)
ITT	interface transparency thesis	(Hypothesis 4:12)
IPM	integrated processing model	(Section 5.3)
LLR	log likelihood ratio	(Section 2.5.2)
LRT	likelihood ratio test	(Definition 2:29)
LF	logical form	
MEG	magnetoencephalography	

mITT	modified ITT	(Hypothesis 4:20)
NPI	negative polarity item	
PDA	pushdown automaton	(Definition 2:16)
PCT	P-cognition thesis	(Hypothesis 3:2)
QUD	question under discussion	
QR	quantifier raising	
RT	reaction time	
ROI	region of interest	
SAT	speed-accuracy tradeoff	
SPRT	sequential probability ratio test	(Section 2.5.2)
SDT	signal detection theory	(Section 2.5)
SMH	strongest meaning hypothesis	(Hypothesis 3:9)
TM	Turing machine	(Definition 2:17)
UE	upward entailing	(Definition 2:8)

1

INTRODUCTION

Sentences that contain quantifiers differ greatly in how difficult they are to process. To illustrate this basic phenomenon, let us focus on one type of example that demonstrates this very clearly. The type of example I have in mind was discussed by Geurts and van der Slik (2005), who considered sentences like in 1:1. Think of these as a description of a local tennis tournament in which the staff of a hospital participated.

(1:1) Every nurse played against more than two doctors.
All doctors are democrats.
Therefore, every nurse played against more than two democrats.

Most people easily recognize that the last sentence follows from the other two. Interestingly, changing one word not only renders the conclusion invalid but also drastically affects the perceived difficulty:

(1:2) Every nurse played against fewer than two doctors.
All doctors are democrats.
Therefore, every nurse played against fewer than two democrats.

Judging the validity of this inference is perceived to be quite difficult. One may ask what causes this increased perceived difficulty. The two words *fewer* and *more* do not differ in syntactic category and thus syntactic factors can be safely excluded. Furthermore, it is difficult to imagine that phonological features or features like lexical frequency are responsible. There are also no obvious pragmatic differences. Finally, the invalidity of the inference in 1:2, as compared to the validity of 1:1, does also not provide an explanation: If we additionally change the quantifier *every* to *at most three*, the conclusion is valid again, but the inference seems to be even more demanding. What remains as an explanation of the difference in perceived difficulty is the lexical semantics of these words and how it interacts with sentence meaning.

The increased difficulty of 1:1 vs. 1:2 is not only accessible through introspection. Actually, Geurts and van der Slik (2005) asked participants in an experiment to judge whether inferences like these are valid. The proportion of erroneous responses was substantially higher in experimental conditions that presented sentences like 1:2 than in conditions that presented sentences like 1:1. This shows that 'semantic difficulty' can be measured. In fact, the semantic difficulty of quantified sentences was investigated experimentally using psycholinguistic methods at least since Just and Carpenter (1971) and especially received attention in recent years. It has been shown that a number of different experimental measurements are sensitive to it. Measurable semantic processing difficulty opens up the potential to study many interesting aspects of sentence meaning and how it is processed. At the same time, classical semantic theory does by itself not provide the tools to describe, explain, or predict semantic processing difficulty.

The present thesis is concerned with the question whether and how we can amend semantic theory in order to model the processing difficulty of quantified sentences. The method is to make minimal and well-motivated processing assumptions, conjoin them to existing semantic theory and to test whether this makes correct predictions. The processing assumptions are motivated from neighboring fields within the cognitive sciences such as computer science or cognitive psychology.

In the rest of this introduction, the relation between theory and experiment in semantics is discussed briefly (section 1.1), the phenomenon under investigation in this dissertation is delimited (section 1.2) and a summary is given (section 1.3)

1.1 THE RELATION BETWEEN THEORY AND EXPERIMENT IN SEMANTICS

Formal semanticists have accumulated insights about how grammatical operations derive the meanings of larger linguistic constructions from the meanings of their lexical parts. Traditionally, the typical data source for semantic theorizing are introspective judgments about sentences or discourse fragments – in particular, about their truth conditions, their possible readings, their semantic well-formedness or entailments among them. Today, the introspective data are complemented by a growing amount of experimental data. Among these are

proportions of response categories or reaction times (RTs) collected in forced-choice tasks, reading time measures obtained from self-paced reading or eye-tracking experiments and also data from EEG, MEG or fMRI studies. Some of these methods have only recently become broadly available.

It has been recognized by semanticists that these kinds of data could potentially be useful to test or develop semantic theory (see e.g. the handbook articles of Bott, Featherston, Radó, & Stolterfoht, 2011 and Krifka, 2011). In addition, semanticists have recently started to ask whether their theories are cognitively plausible (among many others see van Lambalgen & Hamm, 2005; McMillan, Clark, Moore, Devita, & Grossman, 2005; Baggio & van Lambalgen, 2007; Pietroski, Lidz, Hunter, & Halberda, 2009; Bott, 2010; Szymanik & Zajenkowski, 2010; Pylkkänen, Brennan, & Bemis, 2011; Hackl, Koster-Hale, & Varvoutis, 2012). This development may not seem surprising. After all, the classical way to evaluate semantic theory is by comparing its predictions to speaker judgments which are obviously the result of some cognitive process. Thus, to have some understanding of the process that underlies speaker judgments, at some level of abstraction, may even be a necessary condition to develop a precise and accurate semantic theory. Moreover, devising cognitively realistic semantic theories may allow us to make better use of experimental methods.

However, this development is surely not endorsed in every strand of semantics and there is certainly room for disagreement. On the conceptual level, recent developments may thus call for a reevaluation of what semantic theory should be able to describe, explain or predict. A recent discussion of this fundamental question can be found in the target article by Hamm, Kamp, and Van Lambalgen (2006) and responses to it. It may also be worthwhile to take a look at debates in syntax, where analogous questions have been and still are discussed in terms of the competence-performance distinction (Chomsky, 1965). Moreover, the questions whether grammar and language processing constitute two separate systems (for a recent discussion, see Lewis & Phillips, 2015) or what the relation is between linguistics and neuroscience (e.g. Poeppel & Embick, 2005) are also closely related. These fundamental questions are not touched upon here. Instead, it is reflected on what conclusions theoretical semanticists can draw from the growing body of experimental data and how semantic theories could be amended to accommodate the data.

1.1.1 *Conclusions about semantic theory from experiments*

A general problem pertaining to the relation between theory and experiment in semantics is that semantic theories only rarely make falsifiable predictions about the above-mentioned dependent measures. It is rather obvious that classical formal semantic theories (e.g. Kamp & Reyle, 1993; I. Heim & Kratzer, 1998) do not make immediate predictions about, e.g., the time needed to perform a meta-linguistic judgment or to read a word and integrate it into the unfolding interpretation of a sentence. Similarly, electrophysiological data or neuroimaging data are also not predicted by formal semantic theory. Often, the reason for the lack of immediate predictions can be found in the fact that either no processes are described or the processes are not intended as cognitive models. In consequence, it is difficult to evaluate such theories against data of the mentioned kind. Without resorting to additional processing assumptions, hardly any conclusions about the validity of the theories can be drawn from the data.

Proportions of response categories in forced-choice tasks may seem less problematic at first sight because these data resemble the classical introspective judgments rather closely. For example, consider sentence-picture verification. It may seem trivial to derive predictions about the outcome of verification experiments from semantic theory because they should follow from the truth conditions of a sentence – the prime object of study in formal semantics. However, because of the ubiquity of gradience, even this case is problematic. Formal semantic theories generally either predict that a sentence, under one particular reading, is true or that it is false in a particular context. What is, instead, typically found in sentence-picture verification and other forced-choice tasks are graded proportions of judgments. The observed proportions lie anywhere between 0 and 1 and, crucially, are also affected by multiple, possibly non-linguistic factors in a gradual fashion. Most formal semantic theories do not provide the means to predict these kinds of experimental outcomes (for related discussions concerning theories of syntax and grammaticality judgments see, e.g., Sorace & Keller, 2005; Featherston, 2007; Bader & Häussler, 2010).

Strictly speaking, the ubiquity of graded proportions, therefore, leaves only two choices. Firstly, we could reject the classical theories because they are not equipped to describe the experimental outcomes accurately. Secondly, we could reject the idea that they make predictions about the experimental outcomes and thus ignore the ex-

periments as far as semantic theorizing is concerned. For example, we may hold that the graded proportions come about because of extra-linguistic cognitive processes distinct from what semantic theory aims to describe (e.g. core grammatical knowledge or a part of an encapsulated language module; cf. Fodor, 1986). It seems that the latter option is often preferred.

1.1.2 *Accommodating experimental results in semantic theory*

We could make better use of experimental data if we were able to predict the experimental outcomes based on semantic theory. One way to approach this is to introduce additional assumptions and conjoin them to the existing theories. Pursuing this route is attractive as it allows us to maintain the strengths of the existing theories, for example, their compositionality. In fact, this approach is not uncommon, as illustrated by the following highly selective list of examples from the literature.

For instance, to accommodate graded proportions of judgments in forced choice tasks we may introduce uncertainty into the semantic representations (Goodman & Lassiter, 2015). In particular, uncertainty may be introduced at the level of ambiguity resolution (e.g. Brasoveanu & Dotlačil, 2013) or it may be due to the perception of (e.g. Pietroski et al., 2009; Lidz, Pietroski, Halberda, & Hunter, 2011; Scontras, Graff, & Goodman, 2012; Tomaszewicz, 2013) or prior expectations about (Schöller & Franke, 2015, 2016) the non-linguistic context. Another example is to predict RT in forced choice tasks on the basis of some measure of how complex the semantic representations are – the simplest such measure probably being the number of symbols (e.g. H. Clark & Chase, 1972; Geurts, Katsos, Moons, & Noordman, 2010). More symbols may lead to longer response latencies, possibly because the symbols correspond to computational steps.[1] Moreover, in order to predict reading times, we may assume that ambiguity decreases reading speed, especially if there is close competition between readings (e.g. Filik, Paterson, & Liversedge, 2004; Bott & Radó, 2009). Similarly, we may assume that unlikely semantic interpretations take longer to process than likely ones (Brasoveanu & Dotlačil, 2013) or that revision or reanalysis of semantic representations takes time (e.g.

[1] Number of symbols is clearly only one possibility and it does not necessarily correlate with processing complexity (cf. the remark on *expression complexity* and *data complexity* by Szymanik, 2016, p. 104)

Hackl et al., 2012; Bott & Schlotterbeck, 2015).[2] As a final example, in order to predict neural activity based on classical semantic theory, we may assume that different types of semantic composition are carried out in different brain areas (Westerlund, Kastner, Al Kaabi, & Pylkkänen, 2015).

Once experimental results are accommodated in semantic theory via auxiliary processing assumption, we may wish to decide between different alternatives. A major methodological difficulty in testing semantic theories that are amended with auxiliary processing assumptions is that we do generally not know whether a given experimental outcome would falsify the semantic *core theory* or the additional processing assumptions (for a recent discussion of this issue with regard to experimental pragmatics, see Chemla & Singh, 2014a, 2014b).

As an illustration, consider the discussion surrounding one of the the mentioned examples, namely the study of Hackl et al. (2012). The authors discuss experimental data including reading times and offline ratings of sentences involving *antecedent contained deletion*. The purpose of their study was to decide between two competing types of theories on quantifier interpretation, namely *quantifier raising* (QR) and *type shifting* theories. They claimed that their data are compatible with QR but inconsistent with type-shifting theories. In response, Gibson et al. (2015) argued that Hackl et al. made processing assumptions with regard to the QR theory that are implausibly strong. According to Gibson et al., the experimental findings are in fact inconsistent with the QR theory, but can be accounted for by the type shifting theory using yet other processing assumptions. Also related is a comment of Szabolcsi (2014), who argued that, under a broader conception of type-shifting theories, Hackl et al.'s original data are actually just as compatible with type-shifting as they are with QR. Even omitting its details, this discussion exemplifies the difficulty of evaluating a semantic core theory using experimental data in combination with processing assumptions.

From a practical point of view, we should therefore strive to meet two criteria when trying to model experimental data based on formal semantic theory. Firstly, core theories should be preferred that lend themselves to being developed into processing models. This entails, for example, that core theories should not only be formulated

[2] One complication with this kind of auxiliary assumption is that, in formal theories, semantic interpretation is often not described in an incremental fashion but language comprehension often proceeds incrementally. For discussion see Bott and Sternefeld (2017), who discuss challenging cases and propose an incremental event semantics.

explicitly and precisely but also that they should describe efficiently computable functions (Frixione, 2001). Secondly, processing assumptions used to amend semantic core theories should be parsimonious and well-motivated. Optimally, auxiliary processing assumptions are independently motivated from empirical research or through careful theoretical consideration. In the absence of further information, these criteria increase the credibility of a theory and its chance to be compatible with a wide range of data. These criteria will be taken into account in what follows.

Conceptually, a procedural perspective on meaning is taken in the present work. The meaning of a sentence is conceived of as a procedure or an algorithm (cf. the formalism of Moschovakis, 1994) that computes its truth value in context. This approach seems to be suitable for the investigation of semantic processing difficulty, especially since much of the dissertation is concerned with sentence-picture verification. It is surely an unusual approach to semantics but one that is not new and can be traced back to one of the founding fathers of the systematic and formal study of meaning, Gottlob Frege. Recent summaries and discussions of the philosophical position that meanings are procedures can be found in Szymanik (2009, 2016) who refers to Tichý, van Benthem and Suppes among others (see also Pietroski et al., 2009). Here, we leave it at these hints because the philosophical issues are not our main concerns.

1.2 DELIMITING THE PHENOMENON

The present work is concerned with the question whether and how processing difficulty of quantified sentences can be modeled based on semantic theory. The focus on quantified sentences is motivated by the fact that quantifiers can quite confidently be said to be the most studied type of expressions in formal semantics. This implies, among other things, that there is a rich theoretical background we can make use of. In addition, it will become clear as we go that the processing of quantifiers has, especially in the last fifteen years, received a lot of attention and our knowledge about it has grown continuously. The present work contributes to this vibrant development.

The processing of quantified sentences is a complex multi-faceted phenomenon. The focus of the present work is on a small aspect of it: We focus on comprehension and truth evaluation of sentences that contain one quantifier, the main focus being on truth evaluation. This

is justified by the fact that semantic theory is most explicit with regard to what truth conditions sentences or discourse fragments have. Furthermore, truth conditions may be viewed as the one common denominator among semantic theories, which come in many different flavors but share the common goal to describe the truth conditions of sentences in a systematic fashion. Moreover, collecting truth-value judgments is arguably among the empirical methods that are most informative about different aspects of sentence meaning including how it is processed (see e.g. Krifka, 2011, for discussion). Hopefully, studying truth evaluation first will lay the groundwork to study other aspects of the processing of sentence meaning in future research.

A number of topics should be mentioned here that are interesting and important aspects of the processing of quantified sentences but are nevertheless not discussed in any detail in this dissertation. Firstly, the present work is concerned with *semantic* processing and, thus, other aspects of the processing of quantified sentences are not covered. Most notably, pragmatic aspects are neglected. Throughout the dissertation examples, discussions, experimental procedures and experimental materials are chosen or conducted in such a way that relevance of pragmatic factors should be minimal.

Furthermore, we focus on *sentence* meaning and, therefore, discourse related phenomena are not covered. There is obviously a vast literature on discourse related aspects of the interpretation of quantifiers, not only in semantics but also in psycholinguistics. These and related issues are not discussed here. Moreover, reasoning with quantified sentences (in the syllogistic sense) is also not discussed. This aspect of the semantics of quantifiers has, of course, received much attention in linguistics, logic, and psychology and there are many unresolved questions. But it is not discussed here.

Finally, the interpretation of multiply quantified sentences is a topic that has attracted attention by semanticists since the birth of formal semantics and has also received attention in psycholinguistics in more recent years. The processing of multiply quantified sentences, and in particular the resolution of quantifier scope ambiguities is not the subject of this work.

1.3 SUMMARY OF THE DISSERTATION

Chapter 2 introduces basic theoretical concepts that are used throughout the rest of the dissertation.

In chapter 3, processing difficulty is approached on an abstract level. The difficulty of the truth evaluation of quantified sentences is classified using computational complexity theory. Two experiments are reported which tested whether cognitive capacities are limited to those functions that are computationally tractable (PTIME-cognition thesis). In particular, the semantic processing of reciprocal sentences with generalized quantifiers, i.e., sentences of the form *Q dots are directly connected to each other*, where *Q* stands for a generalized quantifier, e.g. *all* or *most* is investigated. Sentences of this type are notoriously ambiguous and it has been claimed that the logically strongest reading is preferred (strongest meaning hypothesis). Depending on the quantifier, the verification of their strongest interpretations may be computationally intractable whereas the verification of the weaker readings is always tractable. A picture completion experiment and a picture verification experiment are reported that investigated whether comprehenders shift from an intractable reading to a tractable reading which should be dispreferred according to the strongest meaning hypothesis. The results from the picture completion experiment indicate that intractable readings occur in language comprehension. Their verification, however, rapidly exceeds cognitive capacities in case the verification problem cannot be solved using simple heuristics. In particular, it is argued that during verification, guessing strategies are used to reduce computational complexity.

Chapter 4 discusses two common approaches to model the canonical verification procedures associated with quantificational sentences. The first is based on the semantic automata model. This model conceives of quantifiers as decision problems or, equivalently, formal languages and characterizes the computational resources that are needed to compute them. The second approach is based on the interface transparency thesis, which stipulates a transparent interface between semantic representations and how verification procedures are realized in the general cognitive architecture. Predictions of both approaches are compared to experimental data from psycholinguistic experiments.

Chapter 5 focuses on a particularly interesting test case that is challenging for both of these approaches. In particular, increased processing difficulty of modified numerals of the form *more than n* as compared to *fewer than n* is investigated. An integrated processing model is proposed for these cases which integrates insights from formal semantics with models from cognitive psychology of how nu-

merical information is represented and processed. This model can be seen as implementation and extension of the interface transparency thesis. The truth evaluation process is conceived of as a stochastic process as decribed in sequential sampling models of decision making. The increased difficulty of *fewer than n* as compared to *more than n* is attributed to an extra processing step of scale-reversal that precedes the actual decision process. Predictions of the integrated processing model are tested and confirmed in two sentence-picture verification experiments. The first is an ordinary sentence-picture verification experiment; the second employs the response-signal speed-accuracy tradeoff procedure in order to control the processing time participants use.

Chapter 6 discusses whether and how the integrated processing model can be extended to other quantifiers. It is shown how it can be extended to proportional comparative quantifiers, like *fewer than half* and *more than half*. This extension is discussed in the light of existing experimental data. Moreover, it is shown that what are called empty-set effects can be naturally derived from the model.

Chapter 7 presents data from two eye tracking experiments that show that *fewer than n* leads to increased difficulty as compared to *more than n* already during reading. The same holds *fewer than half* as compared to *more than half*, especially if these quantifiers are combined with sentence negation. Potential accounts of these findings are discussed and it is speculated how processing difficulty during reading and during verification may be related to each other.

Conclusions are presented in chapter 8.

2

THEORETICAL PRELIMINARIES

The present chapter introduces some theoretical preliminaries that are used throughout the dissertation. The intention in writing the present chapter was, of course, to pull some basic stuff out of the later chapters in order to keep discussions to the point. The reader is neither expected nor supposed to read this chapter from A to Z, especially not on first pass. Rather he or she is invited to skip this chapter and jump back as needed. References are provided throughout the text, so the reader knows when to jump back.

It is presumed that readers are familiar with common formal approaches to the semantics of natural language (for an introduction see, e.g., Gamut, 1991a, 1991b; I. Heim & Kratzer, 1998) and with mathematical logic (e.g. Rautenberg, 2009). Moreover, familiarity with probability theory and statistics is presupposed (e.g. Dekking, 2005). Finally, basic concepts from cognitive psychology (e.g. Anderson, 2015) and neuroscience (e.g. Dayan & Abbott, 2001; Mallot, 2013) are also presupposed.

2.1 TWO PERSPECTIVES ON THEORY BUILDING IN COGNITIVE SCIENCE

Let us start with two influential perspectives on theory building in cognitive science that are useful to have as background for the chapters to follow: the so-called *tri-level hypothesis* of Marr (1982) and the *rational analysis* of Anderson (1990). They are briefly introduced here. Both Marr and Anderson (1990) discuss different "levels of analysis." The interested reader may consult Anderson (1990, pp. 3–23) for a comparison of his own perspective to related discussions in the literature (including, for example, Marr's three levels and the competence-performance distinction of Chomsky, 1965).

2.1.1 *Marr's three levels*

While studying the visual system, Marr (1982) developed his tri-level hypothesis. He proposed that there are three levels at which an "information processing device must be understood before one can be said to have completely understood it" (p. 24):

1. Computational theory:
   What problem does the system solve in terms of an input-output mapping?

2. Representation and algorithmic
   What are the representations of input and output and what is the algorithm that transforms input to output?

3. Hardware implementation
   How is the algorithm that transforms input to output realized physically?

To illustrate these levels, Marr used a cash register. The main purpose of a cash register is to add or subtract prizes of items in an appropriate manner. This is what has to be specified at the first level. In particular, it requires some theory of arithmetic. At the second level, the representations and algorithms are specified. For example, Arabic numerals can be used to represent prices and "for the algorithm we could follow the usual rules about adding the least significant digits first and "carrying" if the sum exceeds 9" (p. 23). The third level describes the physical device in which the algorithm is implemented. Marr stressed that one and the same algorithm may be carried out by many different devices.

Marr noted that, while the three levels may, in principle, be studied individually, there are some dependencies between them. For example, some algorithm may be less efficient but more robust (level 2) than another one. Depending on what problem has to be solved (level 1), one or the other may be better suited. Similarly, there are dependencies between the second and third level. Marr compares the low number of connections in proportion to the number of gates in digital computers to the relatively high number of connections in proportion to nerve cells in a brain. He mentions that the former type of architecture (level 3) is better suited to perform operations in series than in parallel (level 2).

The present work is mostly concerned with the relation between levels one and two. Semantic theory is taken as a level-1 description and it is asked what assumptions need to be added in order to model processing difficulty and how these assumptions are motivated. Usually, it is assumed that an analysis on the second level is needed to model processing difficulty, but see chapter 3 for an approach that derives predictions about processing difficulty more or less directly from the first level and section 4.1 for descriptions that may be said to be situated at level 1.5 (i.e. between levels one and two; cf. Frixione, 2001, p. 382 who refers to Christopher Peacocke). Of course, the study of processing difficulty and the second level of description may also feed back into level-1 theories.

2.1.2 *Andersons's rational analysis*

During his study of the human memory system Anderson (1990) developed a research methodology, called *rational analysis*, that has been very influential in cognitive science: To build a theory of a cognitive function the following steps are iterated (pp. 29–30):

1. Goals:
   Specify precisely the goals of the cognitive system.

2. Environment:
   Develop a formal model of the environment to which the system is adapted.

3. Computational limitations:
   Make minimal assumptions about computational limitations.

4. Optimization:
   Derive the optimal behaviour function given 1–3.

5. Data:
   Examine the empirical evidence to see whether the predictions of the behaviour function are confirmed.

6. Iteration:
   Repeat, iteratively refining the theory.

At several places in this dissertation aspects of this methodology are used. This applies especially to two chapters. The first is chapter

4, where two common approaches are discussed how to characterize canonical verification procedures for natural language quantifiers. The second is chapter 5, where an processing model is proposed that integrates insights from linguistics and cognitive psychology.

2.2 GENERALIZED QUANTIFIERS AND SOME OF THEIR PROPERTIES

The term generalized quantifier (GQ) is of particular importance for the present work (see Peters & Westerståhl, 2006, for an introduction and comprehensive overview). This term has slightly different, but related, usages in mathematical logic and formal linguistics. Here, we provide the general definition due to Per Lindström (cf. Szymanik, 2009) in order to be able to capture the semantics of natural language determiners as well as quantified reciprocals (the topic of chapter 3).

Definition 2:1 (Generalized Quantifier, GQ, Lindström, 1966)**.** *Let $a_1, ..., a_n$ be integers. A generalized quantifier Q of type $t = (a_1, ..., a_n)$ is a class of models (i.e. relational structures) of a vocabulary $\sigma_t = \{R_1, ..., R_n\}$, such that R_i is a_i-ary, for $1 \leq i \leq n$, and Q is closed under isomorphism, i.e. if $\mathcal{M}$ and $\mathcal{M}'$ are isomorphic, then $\mathcal{M} \in Q$ iff $\mathcal{M}' \in Q$.*

As usual, a model $\mathcal{M}$ of a vocabulary $\{R_1, ..., R_n\}$ consists of a non-empty set M, called domain (or universe) of $\mathcal{M}$, and relations $R_i^{\mathcal{M}} \subseteq M^{a_i}$ for each predicate symbol R_i from the vocabulary, i.e. $\mathcal{M} = (M, R_1^{\mathcal{M}}, ..., R_n^{\mathcal{M}})$. Note that, a generalized quantifier Q characterizes a second order relation Q_M between first order relations over a domain M. We may thus also write $Q_M(R_1^{\mathcal{M}}, ..., R_n^{\mathcal{M}})$ instead of $\mathcal{M} \in Q$. In the following, the superscript is ommitted and relations are referred to with predicate symbols since the context provides disambiguation in most cases.

Example 2:2. *The generalized quantifiers $\exists^2$ of type (1) and* AtLeastTwo*,* Every *and* Most *of type $(1,1)$ can be defined as:*

$$\begin{aligned} \exists^2 &:= \{\mathcal{M} : |R_1| \geq 2\} \\ \textsc{AtLeastTwo} &:= \{\mathcal{M} : |R_1 \cap R_2| \geq 2\} \\ \textsc{Every} &:= \{\mathcal{M} : |R_1 \setminus R_2| = 0\} \\ \textsc{Most} &:= \{\mathcal{M} : |R_1 \cap R_2| > |R_1 \setminus R_2|\}. \end{aligned}$$

Corresponding to a generalized quantifier Q, we may introduce a variable binding operator Q in order to build formulas in the following way.[1]

Definition 2:3 (Logics with Generalized Quantifiers)**.** *Let $\mathcal{L}$ be a logic and Q a generalized quantifier of type $(a_1, ..., a_n)$. $\mathcal{L}$ is extended to the logic $\mathcal{L}(Q)$ via the following two rules:*

- *If $\overline{v_1}, ..., \overline{v_n}$ are a_i-tuples of pairwise distinct variables ranging over individuals and $\phi_1[\overline{v_1}], ..., \phi_n[\overline{v_n}]$ are formulas, then the following is a formula:*
$$Q\overline{v_1}, ..., \overline{v_n}\ \left(\phi_1[\overline{v_1}], ..., \phi_n[\overline{v_n}]\right).$$
- *$Q\overline{v_1}, ..., \overline{v_n}\ \left(\phi_1[\overline{v_1}], ..., \phi_n[\overline{v_n}]\right)$ is true in a model $\mathcal{M}$ iff*
$$Q_M\left(\phi_1^{\mathcal{M},\overline{v_1}}, ..., \phi_n^{\mathcal{M},\overline{v_n}}\right),$$
where $\phi_i^{\mathcal{M},\overline{v_i}} = \{\overline{a} \in M^{a_i} : \phi_i[\overline{a}/\overline{v_i}]$ is true in $\mathcal{M}\}$.

Below in chapter 4, we will classify quantifiers according to *definability* and correlate this with the computational resources needed to decide whether $Q_M(A, B)$. For this purpose, the following definition is useful.

Definition 2:4 (Definability of GQs)**.** *Let Q be a GQ of type t and $\mathcal{L}$ a logic. We say Q is definable in $\mathcal{L}$ iff there is an $\mathcal{L}$-sentence ϕ of vocabulary σ_t, such that for every σ_t-structure $\mathcal{M}$: $\mathcal{M} \models \phi \Leftrightarrow \mathcal{M} \in Q$.*[2]

In formal semantics, GQs are often equated with determiner denotations. The reason is that natural language determiners, such as *some*, *all*, or *most*, can be analyzed as GQs of type $(1, 1)$. For example, consider sentences of the form DET *A are B*, where A and B denote unary relations (e.g. nouns). We can analyze the determiner DET as a generalized quantifier Q by demanding that DET *A are B* is true in a model $\mathcal{M}$ iff $Q_M(A, B)$. Equivalently, we may translate DET *A are B* to $\mathsf{Q}x, y\ (Ax, By)$ (by convention also written as $\mathsf{Q}x\ (Ax, Bx)$). To give a simple, concrete example, we may analyze the determiner *all* as denoting the type $(1, 1)$ quantifier $\{\mathcal{M} : A \subseteq B\}$. We call determiners that denote type $(1, 1)$ generalized quantifiers *quantificational*

[1] The term *logic* used in Definition 2:3 and below can simply be understood as the combination of a syntax and a semantics as found in introductions to first order logic (e.g. van Dalen, 1994). Some discussion and references to literature is provided by Peters and Westerståhl (2006, pp. 449–451).

[2] As usual, $\mathcal{M} \models \phi$ means that formula ϕ is true in model $\mathcal{M}$ (see e.g. Rautenberg, 2009).

determiners. Moreover, the first argument of a quantificational determiner is referred to as its *restriction* (or restrictor) and the second is referred to as its *scope*.

Note on terminology. *The term "generalized quantifier theory" is ambiguous. In linguistic contexts, it is often used to refer to a collection of hypotheses about the semantics of determiners, determiner phrases or noun phrases. For example, it may be used to refer to a collection of hypotheses put forward by Barwise and Cooper (1981) or in related work. The term is not used this way here, but simply refers to the theory of GQs. The aforementioned type of hypotheses are considered linguistic applications of generalized quantifier theory (GQT).*

The mentioned approach to the semantics of determiners has produced a fruitful line of research within formal semantics (see e.g. Keenan, 2006; Peters & Westerståhl, 2006, for overview). One achievement of this research was to differentiate quantificational determiners with respect to their logical properties and formulate grammatical generalizations based on these properties. One famous example concerns the distribution of *negative polarity items* (NPIs) like *anything* or *ever*. After observing that certain determiners (or noun phrases) *license* NPIs (examples 2:5-a,d) whereas others do not (2:5-b,c,e,f), one may ask which exactly the licensors are and what they have in common (apart from being an NPI licensor). A good first approximation can be given in terms of a property called *monotonicity*: NPIs occur only in *downward monotone* contexts (Fauconnier, 1978; Ladusaw, 1980, but see also Giannakidou, 2011 for a refined hypothesis).

(2:5) a. No student had read anything about gauge theory.
b. *Every student had read anything about gauge theory.
c. *Some student had read anything about gauge theory.
d. No student had ever read about gauge theory.
e. *Every student had ever read about gauge theory.
f. *Some student had ever read about gauge theory.
(modfied from Ladusaw, 1980)

Below, *monotonicity* will be of relevance. This property depends on the inferences that a particular quantifier allows. To get the idea across, we can say that *monotone increasing* quantifiers (or operators in general) are those that license inferences from more specific to less specific expressions (cf. 2:6), whereas *monotone decreasing* quantifiers

license inferences from less specific to more specific expressions (cf. 2:7). A formal definition is given below.

(2:6) Every student sings and dances.

⇒ Every student sings.

↛ Every student sings, dances and claps hands.

(2:7) No student sings and dances.

↛ No student sings.

⇒ No student sings, dances and claps hands.

Definition 2:8 (Monotonicity)**.** *A generalized quantifier Q of type $(a_1, ..., a_n)$ is called:*

- *monotone increasing in its i-th argument iff from $Q_M[R_1, \ldots, R_n]$ and $R_i \subseteq R_i' \subseteq M^{a_i}$ follows $Q_M[R_1, \ldots, R_{i-1}, R_i', R_{i+1}, \ldots, R_n]$,*
- *monotone decreasing in its i-th argument iff from $Q_M[R_1, \ldots, R_n]$ and $R_i' \subseteq R_i \subseteq M^{a_i}$ follows $Q_M[R_1, \ldots, R_{i-1}, R_i', R_{i+1}, \ldots R_n]$*
- *non-monotone in its i-th argument iff it is neither monotone increasing nor decreasing in that argument.*

Convention. *If a generalized quantifier of type $(1,1)$ is monotone increasing or decreasing in its second argument, we say that it is upward entailing (UE) or downward entailing (DE), respectively. Otherwise, we call it non-monotone.*

2.2.1 *CE-quantifiers as relations between numbers*

Other properties of quantifiers have been hypothesized to be *semantic universals*, i.e. properties that all quantificational determiners in every natural language have (Barwise & Cooper, 1981; Keenan & Stavi, 1986; van Benthem, 1986), and are still considered as such. Two of these properties are relevant here. As we will see shortly, these two properties in combination lead to yet another conception of quantificational determiners, namely as relations between numbers. The first is *domain independence* (or EXT which is mnemonic for *extension*).

Definition 2:9 (Domain Independence, EXT)**.** *A quantifier Q of type $(1,1)$ is domain independent iff $Q_M(A,B) \Leftrightarrow Q_{M'}(A,B)$, for $A, B, \subseteq M \subseteq M'$.*

EXT has the effect that no entities outside the set $A \cup B$ matter if we want to decide whether $Q_M(A,B)$ holds. The second property

is *conservativity* (CONS), which reduces the relevant elements even further. If a quantifier is conservative we do not need to consider any elements that are in B but not in A.

Definition 2:10 (Conservativity[3], CONS)**.** *A quantifier Q of type $(1,1)$ is called conservative iff $Q_M(A,B) \Leftrightarrow Q_M(A, A \cap B)$, for $A, B \subseteq M$.*

GQs of type $(1,1)$ that satisfy both, CONS and EXT are called *CE-quantifiers*. Both properties in combination have the effect that only the sets $A \cap B$ and $A \setminus B$ are relevant to decide whether $Q_M(A,B)$. Moreover, because GQs are closed under isomorphisms, only the cardinalities $|A \cap B|$ and $|A \setminus B|$ count. As a consequence CE-quantifiers can be considered two-place relations between cardinal numbers.

Proposition 2:11 (CE-quantifiers as relations over cardinal numbers)**.** *A quantifier Q is a CE-quantifier iff there is a (unique) binary relation R_Q over cardinal numbers such that $R_Q(|A \setminus B|, |A \cap B|) \Leftrightarrow Q_M(A,B)$, for $A, B \subseteq M$.*

Proof. Follows from Peters and Westerståhl (2006, pp. 158-160). □

The following corollary tells us that CE-quantifiers can be identified with subsets of $\mathbb{N}^2$ once we restrict ourselves to finite structures.

Corrolary 2:12. *A quantifier Q is a CE-quantifier iff there is a (unique) binary relation R_Q over $\mathbb{N}$ such that, for any finite sets M and $X, Y \subseteq M$,*

$$R_Q(|X \setminus Y|, |X \cap Y|) \Leftrightarrow Q_M(X,Y).$$

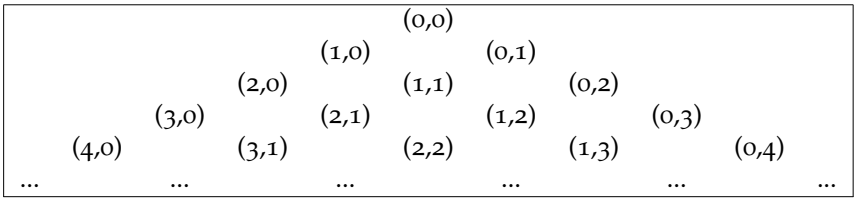
***Example* 2:13.** *Here are the relations for the CE-quantifiers from above:*

$$\begin{aligned} R_{\text{AtLeastTwo}}(a,b) &\Leftrightarrow b \geq 2 \\ R_{\text{Every}}(a,b) &\Leftrightarrow a = 0 \\ R_{\text{Most}}(a,b) &\Leftrightarrow b \geq a. \end{aligned}$$

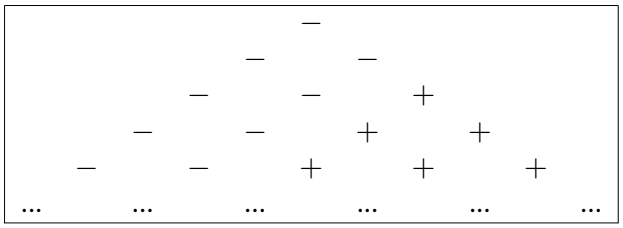
Note. *In what follows, we restrict ourselves to finite structures.*

For illustration of the relations just introduced, the so-called number tree is useful (van Benthem, 1986). In the number tree, each node corresponds to some $(x,y) \in \mathbb{N}^2$. This is shown in Figure Ia. To represent a CE-quantifier Q, we replace (x,y) with the symbol $+$ whenever $R_Q(x,y)$ holds and with $-$ otherwise. Examples are shown in Figure

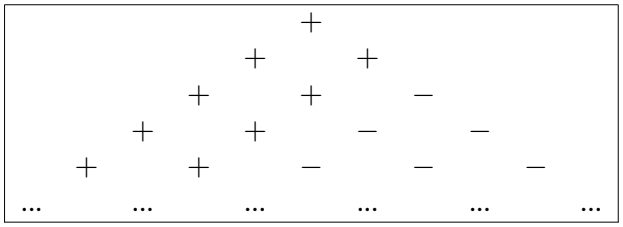
[3] A completely parallel definition applies to the concept of quantifiers used by Keenan and Stavi (1986), which is discussed in section 6.1.



(a) General format



(b) AtLeastTwo



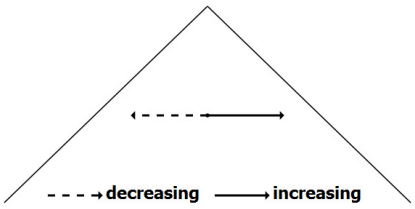
(c) AtMostTwo

Figure I. CE-quantifiers in the number tree.

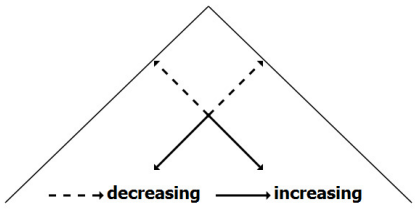
Ib and Ic. These are, of course, just geometrical representations of the characteristic function of R_Q. They are useful because they reveal certain properties of CE-quantifiers. For example, UE quantifiers have the property that pluses propagate to the right, so to speak (see Figure IIa). Moreover, in the number tree representations of quantifiers that are monotone increasing or decreasing in their first argument, pluses propagate either upward or downward along the direction of the two 'axes' of the number tree (see Figure IIb) and quantifiers that are *intersective*, which means that $Q_M(A, B)$ is equivalent to $Q_M(B, A)$, have the property illustrated in Figure IIc.

2.3 DECISION PROBLEMS AND AUTOMATA

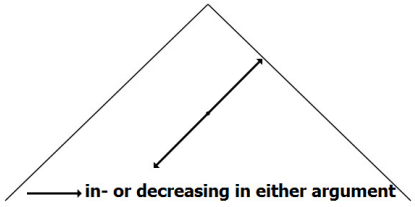
A large part of this dissertation is concerned with truth-value judgments for quantified sentences. Therefore, the concept of a decision problem is useful. A decision problem essentially is a function that maps some input to 1 or 0. Thus, truth-value judgments can be considered a special case. This perspective also opens up connections to



(a) Mononicity 2nd argument



(b) Mononicity 1st argument



(c) Intersective quantifiers

Figure II. Some properties of CE-quantifiers in the number tree. Pluses propagate in the direction of the arrows.

the theory of formal languages because decision problems and formal languages are precisely the same thing. The following definition is taken from Arora and Barak (2009). As usual, the notation Σ^* refers to the *Kleene closure* of Σ (see Arora & Barak, 2009, p. 1 and see also Hopcroft & Ullman, 1979 for an overview of formal languages and automata theory).

Definition 2:14 (Decision problem)**.** *Let Σ be any finite set, called an alphabet. A decision problem is a function $f : \Sigma^* \to \{0,1\}$. We identify f with the subset $L_f := \{x \in \Sigma^* : f(x) = 1\}$ of Σ^* and also refer to L_f as a formal language.*

Languages (or decision problems) can be categorized according to the Chomsky hierarchy and also according to the type of computational device needed to recognize (or compute) them (see e.g. Hopcroft & Ullman, 1979). A relatively simple class of languages are the *regular languages*. There are several equivalent ways to define regular languages. One example, is to define the type of grammar that generates these languages, called *regular grammars*. Another possibility, is to determine which kind of computational device is needed to recognize them. The regular languages are exactly those languages that are recognized by a certain type of computing device, namely a *deterministic finite state automaton (DFA)*. This means that every regular language can be associated with a DFA that computes the corresponding function, and *vice versa*.

Definition 2:15 (Deterministic finite state automaton)**.** *A deterministic finite state automaton is a quintuple* $(Q, \Sigma, \delta, q_0, F)$*, where:*

- Q *is a finite set of states;*
- Σ *is an alphabet called the input alphabet;*
- $\delta : Q \times \Sigma \to Q$ *is a function called transition function;*
- $F \subseteq Q$ *is the set of accepting states;*
- $q_0 \in Q$ *the start state.*

A DFA works as follows: At the beginning of a computation, it is in its initial state q_0. Then, it reads input symbols one after the other and changes its current state according to the transition function δ. A picture helps to illustrate the workings of this kind of device. Consider the simple regular language $1^* := \{\epsilon, 1, 11, \ldots\}$, which contains the empty string, ϵ, and any string of only 1s. Figure IIIa shows a DFA that recognizes this language. It starts in the initial state and then behaves as can be read off from the labeled arrows: Every time it reads a 1 it follows the loop and does not change state. As soon as a 0 is encountered it moves to q_1. Once it is there, no arrow leads back and it has to stay. The state q_0 is an accepting state, as indicated by the double circle. If no symbol is left to read and the automaton is in q_0, it accepts the string, otherwise it rejects it.

A well-known non-regular language is $\{1^n 0^n : n \geq 1\}$. This is the language that consists of strings with a certain number of 1s followed by the same number of 0s. Another example is the language $\{s \in \Sigma^* : \#_1(s) > \#_0(s)\}$, the subset of $\{0,1\}^*$ that contains a larger number of 1s than 0s ($\#_1(s)$ denotes the number of 1s in s, $\#_0(s)$ the number of 0s). Both of these languages are called *context-free*. To show that these languages are not regular and thus not recognized by a DFA, the *pumping lemma* may be used (see e.g. Hopcroft & Ullman, 1979, pp. 55-56). To recognize languages like these, a more capable device is needed. Both of them are recognized by *pushdown automata (PDA)*.

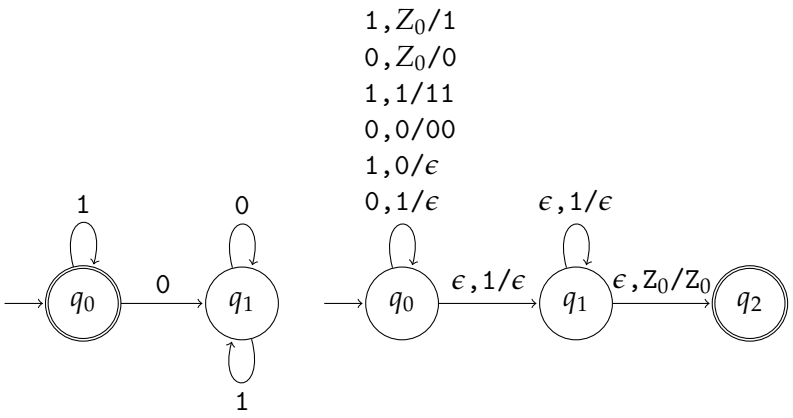
Definition 2:16 (Pushdown automaton, PDA)**.** *A (non-deterministic) PDA is a tuple* $(Q, \Sigma, \Gamma, \delta, q_0, Z_0, F)$*, where:*

- Q *is a finite set of states;*
- Σ *is an alphabet called the input alphabet;*

- Γ *is an alphabet called the stack alphabet;*
- $q_0 \in Q$ *is the initial state;*
- $Z_0 \notin \Gamma$ *is a stack symbol called start symbol;*
- δ *is a mapping from* $Q \times (\Sigma \cup \{\epsilon\}) \times \Gamma$ *to finite subsets of* $Q \times \Gamma^*$*;*
- $F \subseteq Q$ *the set of accepting states.*

A PDA makes use of a memory device called a *stack*. This is a *'last in, first out'* data type. Symbols can be *pushed* onto or *popped* off the top of the stack. The mapping δ determines the permissible "moves" given the current state, input symbol and top symbol on the stack. A string is accepted by PDA if reading the string and following δ either permits an accepting state to be reached or emptying the stack. A PDA that recognizes the language $\{s \in \Sigma^* : \#_1(s) > \#_0(s)\}$ is shown in Figure IIIb. This automaton works similar to the DFA to its left. The main difference is that each state transition is accompanied by a stack operation. The arrow labels are read as follows: $q, z/\gamma$ means that the transition may be followed if the input symbol is q and z is the top stack symbol. At the same time, z is replaced with γ. The automaton in Figure IIIb pushes 1s and 0s onto the stack (top four lines above the loop over q_0) until a pair of a 1 and a 0 is encountered (one on the stack, one in the input; bottom two lines above q_0). In the latter case, the top stack symbol is popped and the next input symbol is read. This way, 1s and 0s are matched in number. An input string is accepted if only 1s but no 0s are left on the stack in the end.

A language that is not context-free is, for example, $\{a^n b^n c^n : n \in \mathbb{N}\}$. This can be shown using a pumping lemma for context-free languages (Hopcroft & Ullman, 1979). This language belongs to the class of *context-sensitive* languages. Because context-sensitive languages are not of much importance in what follows, we skip one level of the Chomsky hierarchy and introduce the Turing machine next. This kind of device recognizes the *recursively enumerable* languages and is thus the mightiest device in the Chomsky hierarchy. I provide the definition in which the machine has k tapes (Arora & Barak, 2009) here. A tape is devided into cells that may each contain a symbol. There is no restriction on the length of a tape. The first tape is called the *input tape* and can only be read; the other $k-1$ tapes are called *work tapes* and are read and write tapes; the last of these is a designated output tape. The symbols L, S and R that occur in the definition below deter-



(a) DFA that recognizes 1^* *(b)* PDA that recognizes $\{s : \#_1(s) > \#_0(s)\}$

Figure III. Two example automata

mine whether the k read-write heads move left, stay or move right at the end of each computational step.

Definition 2:17 (Deterministic Turing machine)**.** *A (k-tape) Turing machine is a tuple* (Γ, Q, δ)*, where:*

- Γ *is a set of symbols, including a designated 'blank' symbol* $\square$*, a designated 'start' symbol* $\triangleright$ *and the symbols* 1 *and* 0*;*
- Q *is a set of states, including a 'start' state* q_{start} *and a designated 'halt' state* q_{halt}*;*
- $\delta : Q \times \Gamma^k \to Q \times \Gamma^{k-1} \times \{L, S, R\}^k$ *is called transition function*

I will not give examples of Turing machines (TMs) that compute specific problems. Rather, they are introduced here for the following two reasons. Firstly, the following thesis is generally believed, which makes TMs *the* model of computation.

Hypothesis 2:18 (Church-Turing thesis)**.** *Any function that is computable (in an informal and intuitive sense) can be computed by a* TM*.*

A reason to believe this thesis is that numerous alternative models have been proposed that could be proven to be equivalent to the TM model. Moreover, I am not aware of a counterexample. Secondly, because of their universal character, TMs will be important as a theoretical concept in the next section, which is about Computational Complexity Theory.

2.4 BASIC NOTIONS FROM COMPUTATIONAL COMPLEXITY THEORY

In chapter 3, quantified sentences are distinguished according to their *computational complexity*, the object of study in computational complexity theory (see Arora & Barak, 2009, from whom the present section is adopted, for an introduction and comprehensive overview). The presents section introduces the two important complexity classes **P** and **NP** and also discusses the complexity of two example problems that are relevant in chapter 3.

2.4.1 *Two important complexity classes*

The most basic concept that is needed in order to talk about computational complexity is:

Definition 2:19 (Running time). *Let $f : \{0,1\}^* \to \{0,1\}^*$ and $T : \mathbb{N} \to \mathbb{N}$ be some functions and M a TM. We say M computes f in $T(n)$-time if for every $x \in \{0,1\}^*$, if M is initialized to the start configuration on input x, then after at most $T(|x|)$ steps it halts with $f(x)$ written on its output tape (written as $M(x) = f(x)$).*

Given the definition of running time, we can introduce the class **P** of functions that can be computed in polynomial time by a deterministic TM. Firstly, we define classes of functions according to running times.

Definition 2:20. *Let $T : \mathbb{N} \to \mathbb{N}$ be some function. We denote with $\mathbf{DTIME}(T(n))$ the set of all decision problems that are computable in $c \cdot T(n)$-time for some constant c.*

Secondly, we define **P** as the union of the classes of functions with *polynomial running time*.

Definition 2:21 (The class **P**). $\mathbf{P} := \bigcup_{c \geq 1} \mathbf{DTIME}(n^c)$

Another important complexity class is **NP**, which stand for *non-deterministic polynomial* time. The original definition of this class makes use of the concept of a non-deterministic Turing machine. A useful intuitive way to think of this kind of device is that it simultaneously tries all the potential 'solutions' of a problem. The class **NP** can, however, also be defined on the basis of a deterministic TM. Here, the idea is that a computational problem belongs to the class iff it can be

checked in polynomial time whether a potential solution is correct. With respect to decision problems, 'solutions' are what is called *certificate* in the following definition. Certificates can be thought of as proofs of the correct answer or hints at the correct answer.

Definition 2:22 (The class **NP**). *A decision problem f is in* **NP** *if there exists a polynomial $p : \mathbb{N} \to \mathbb{N}$ and a polynomial-time TM M, such that for every $x \in \{0,1\}^*$,*

$$x \in L_f \Leftrightarrow \exists u \in \{0,1\}^{p(|x|)} : M(x,u) = 1.$$

If $x \in L_f$ and $u \in \{0,1\}^{p(|x|)}$ satisfy $M(x,u) = 1$ then we call u a certificate for x, w.r.t. L and M.

It is obvious that $\mathbf{P} \subseteq \mathbf{NP}$ whereas the question whether $\mathbf{P} = \mathbf{NP}$ is a famous unsolved mathematical problem. An important class of problems among those in **NP** are the *complete* ones:

Definition 2:23 (Reducibility, **NP**-hardness and completeness). *We say that a decision problem f is polynomial-time reducible to a decision problem g, denoted by $f \leq_p g$, if there is a polynomial time computable function $h : \{0,1\}^* \to \{0,1\}^*$ such that for every $x \in \{0,1\}^*$, $f(x) = 1$ iff $g(h(x)) = 1$.*

We call g **NP***-hard if $f \leq_p g$ for every f in* **NP***. If g is, in addition, in* **NP** *it is called* **NP***-complete.*

NP-complete and -hard problems are not uncommon in actual computer applications (see e.g. Garey & Johnson, 1979). These problems are called *intractable* because under the assumption that $\mathbf{P} \neq \mathbf{NP}$ efficient solutions of these problems are beyond reach even if we factor in that actual computing devices become more and more efficient (cf. Arora & Barak, 2009, ch. 1 & 2). **NP**-complete problems are the most difficult problems in **NP** in the sense that, by the reductions mentioned in the above definition, a way to solve one of these problems efficiently also constitutes an efficient solution to all the others (modulo a polynomial time transformation). If one polynomial time algorithm for any of these problems were found, it would have been shown that $\mathbf{P} = \mathbf{NP}$.

2.4.2 *Two problems*

In chapter 3, one **NP**-complete and one **NP**-hard problem are relevant because these correspond to the verification of certain readings of

natural language sentences, in particular what we call *complete graph readings* of quantified reciprocals (cf. Szymanik, 2010). These two problems are described here briefly.

The CLIQUE problem consists in deciding whether there is a completely connected subgraph of a certain size in a graph (see Garey & Johnson, 1979, problem GT19). It is one of the 21 classical **NP**-complete problems presented by (Karp, 1972).

Definition 2:24 (CLIQUE problem)**.** *Given a finite graph G and a positive integer k, is there a complete subgraph of k or more vertices?*

The MAXCLIQUE problem is related to CLIQUE. It consists in deciding whether the maximal connected subgraph in a graph is of a certain size.

Definition 2:25 (MAXCLIQUE problem)**.** *Given a finite graph G and a positive integer k, is the maximum complete subgraph of G of size k?*

Proposition 2:26. *MAXCLIQUE is* **NP***-hard.*

A simple proof of the **NP**-hardness of MAXCLIQUE proceeds via a polynomial-time reduction from 3SAT (see e.g. Garey & Johnson, 1979, p. 46). This reduction is parallel to the one used by Arora and Barak (2009, p. 51) to proof **NP**-completeness of another well-known problem called INDSET (for "independent set"), which is complementary to the CLIQUE problem.

2.5 HYPOTHESIS TESTING AND SIGNAL DETECTION THEORY

In parts of chapters 4 and 5, it is assumed that the cardinality of a set (also referred to as *numerosity*) is mentally represented in a noisy and analog format (Moyer & Landauer, 1967). Under this assumption, performance in tasks that involve comparison of numerosities can be described using signal detection theory (SDT, see e.g. Green & Swets, 1966) and related approaches. Crucially, this also applies to the verification of certain quantifiers, e.g. proportional quantifiers (cf. Pietroski et al., 2009) and, as is argued in chapter 5, also numerical ones.

2.5.1 *Hypothesis testing problems and decision rules*

In order to describe such tasks, we think of them as *hypothesis testing problems*, which are defined as follows (most of the present section is taken from Poor, 1994).

Definition 2:27 (Hypothesis-testing problem)**.** *Let Y be a random variable and P_0 and P_1 two probability distributions. By a hypothesis-testing problem we refer to the problem of deciding which of the following two hypotheses is true:*

$$H_0 : Y \sim P_0$$
$$H_1 : Y \sim P_1.$$

A hypothesis testing problem is solved using a *decision rule*. Such a rule specifies when we choose H_0 and when we choose H_1.

Definition 2:28 (Decision rule)**.** *Let (Γ_0, Γ_1) be a partition of the observation set Γ. A decision rule is a function*

$$\delta : \Gamma \to \{1, 0\}, \qquad y \mapsto \begin{cases} 1, & \text{if } y \in \Gamma_1 \\ 0, & \text{if } y \in \Gamma_0. \end{cases}$$

In what follows, we let $L(y)$ denote $\frac{p_1(y)}{p_0(y)}$, where p_i is the *density* of P_i, for $i \in \{0, 1\}$. We call this quantity the *likelihood ratio*. Moreover, the following type of decision rule is central.

Definition 2:29 (*likelihood ratio test*, LRT)**.** *A decision rule is called likelihood ratio test if, for some threshold τ_ϕ, it has the form:*

$$\delta_\phi(y) := \begin{cases} 1, & \text{if } L(y) \geq \tau_\phi \\ 0, & \text{if } L(y) < \tau_\phi. \end{cases}$$

Decision rules can be designed to fulfill different optimality criteria. One type of optimal decision rules are called *Bayes rules*. They are based on minimization of the so-called *Byes risk*.

Definition 2:30 (Bayes risk)**.** *Assume that, for $i, j \in \{1, 0\}$, $C_{ij} \geq 0$ represents the cost of choosing H_i if H_j is true and assume that we can assign a (prior) probability π_i to the event that H_j is true. Then, the Bayes risk r of a decision rule δ is defined as:*

$$r(\delta) := \sum_{j=0}^{1} \sum_{i=0}^{1} \pi_j C_{ij} P_j(\Gamma_i).$$

Proposition 2:31. *Given any hypothesis testing problem (and assuming $C_{11} < C_{01}$), the following decision rule, called Bayes rule, is optimal in the sense that it minimizes Bayes risk:*

$$\delta_B(y) := \begin{cases} 1, & \text{if } L(y) \geq \tau_B \\ 0, & \text{if } L(y) < \tau_B \end{cases}, \quad \text{where } \tau_B := \frac{\pi_0 C_{10} - C_{00}}{\pi_1 C_{01} - C_{11}}.$$

Proof. See Poor (1994, section II.B). □

The following cost assignment is called *uniform*. Under a uniform cost assignment, τ_B simplifies to $\frac{\pi_0}{\pi_1}$ and it follows as a corollary that, in this case, the Bayes rule also minimizes the overall probability of error.

$$C_{ij} = \begin{cases} 0, & \text{if } i = j \\ 1, & \text{if } i \neq j \end{cases}$$

More generally, in case no cost assignment is available, an optimality criterion can be formulated on the basis of trading off chances of falsely rejecting H_0 (called *false alarm*) against the chances of falsely rejecting H_1 (called *miss*). The *Neyman-Pearson criterion* is to fix an upper bound on the probability of a false alarm and then to minimize the probability of a miss. The *Neyman-Pearson lemma* (see Poor, 1994, pp. 23–25) states that, for a given hypothesis testing problem and a fixed false alarm probability, α, there exists a unique (randomized) LRT, and thus a unique threshold $\tau_{NP,\alpha}$ that minimizes the probability of a miss.

It is obvious that, in case the likelihood ratio, $L(y)$, depends in a strictly monotone fashion on y, a decision rule can be formulated that is equivalent to the LRT but is based on y instead of $L(y)$: Instead of comparing $L(y)$ to τ_ϕ we compare y to $L^{-1}(\tau_\phi)$.

2.5.2 *Sequential hypothesis testing*

The approach of the previous section can be generalized to the case where observations are sampled in sequence until a decision can 'confidently' be made. In the present section, a generalization of the LRT, called the *sequential probability ratio test* (SPRT), is introduced first; after that, some basic stochastic processes are defined; next, it is shown that certain sequential hypothesis testing procedures can be understood in terms of these processes; and finally, it is discussed how all

of this can be used to model decision making processes as studied in psychological experiments. Most of the present section is taken from Poor (1994) and Bogacz, Brown, Moehlis, Holmes, and Cohen (2006), but see also Stone (1960), Ratcliff (1978) and Gold and Shadlen (2001).

2.5.2.1 *The sequential probability ratio test*

Suppose that instead of one realizations, y, our decision is based on a sequence, $(y_i)_{i\in\mathbb{N}}$, of realizations of independent and identically distributed (i.i.d.) random variables, $(Y_i)_{i\in\mathbb{N}}$. The corresponding hypothesis testing problem is:

$$\begin{aligned} H_0 &: Y_k \sim P_0, \qquad k = 1, 2, \ldots \\ H_1 &: Y_k \sim P_1, \qquad k = 1, 2, \ldots. \end{aligned}$$

Because the Y_i are independent, the likelihood ratio after i-th observation is:

$$LR_i = \prod_{1\leq j\leq i} \frac{p_1(y_j)}{p_0(y_j)}. \tag{2:32}$$

The so-called *sequential probability ratio test* (SPRT) is based on two predefined decision boundaries, $\underline{\tau}$ and $\overline{\tau}$. As long as $\underline{\tau} < LR_i < \overline{\tau}$, no decision is made. As soon as one of the decision boundaries is crossed, a decision is made according to the following rule.

$$\delta_S(y) := \begin{cases} 1, & \text{if } LR_i \geq \overline{\tau} \\ 0, & \text{if } LR_i \leq \underline{\tau} \end{cases} \tag{2:33}$$

It can be shown that the SPRT is optimal in the sense that it needs the fewest possible sample observations y_i on average to achieve fixed error probabilities, which are determined by the boundaries. This theorem is due to Wald and Wolfowitz (1948) and is, for example, discussed by Poor (1994, section III.D) and Bogacz et al. (2006). As before, it is possible to formulate specific optimality criteria, such as, for example, minimizing an analogue of the above-mentioned Bayes risk, and choose the decision boundaries accordingly. What such criteria have in common is that they weigh the expected number of observations that is needed to reach a decision against error proba-

bilities. This is, however, beyond the scope of our present discussion (for more details see e.g. Bogacz et al., 2006).

2.5.2.2 *Random walks and diffusion processes*

In the present section, some basic stochastic processes are introduced. These are adopted from (Kannan, 1979) and (Lawler, 1995). In the next section, they are related to hypothesis testing problems.

Definition 2:34 (Random walk)**.** *A family of random variables $(X_i)_{i\in\mathbb{N}}$ is called a time discrete stochastic process. If such a process has the form,*

$$X_i = X_0 + \sum_{i>0} Y_i,$$

and $(Y_i)_{i\in\mathbb{N}^+}$ are i.i.d., then it is called a random walk.

The time continuous analogue of a random walk is a *drift diffusion process*. These processes are generalizations of the so-called *Wiener process*, often denoted by W. Therefore, we define the latter first. In the following definition, $\mathcal{N}$ denotes a normal distribution.

Definition 2:35 (Wiener Process)**.** *A Wiener process is is a family of random variables $(X_i)_{i\in\mathbb{R}^+}$ that satisfies the following conditions.*

(1) $X_0 = 0$ *with probability one.*

(2) if $s_1 \leq t_1, s_2 \leq t_2, \ldots s_n \leq t_n$*, then* $X_{t_1} - X_{s_1}, X_{t_2} - X_{s_2}, \ldots X_{t_n} - X_{s_n}$ *are independent.*

(3) For $0 \leq s < t$*,* $Y_t - Y_s \sim \mathcal{N}(0, (t-s))$*.*

A Wiener process is a special case of a drift diffusion process. It is called stationary since it has zero expected increments, or zero drift. In general, drift diffusion processes need not be stationary and may have non-zero drift.

Definition 2:36 (Drift diffusion process)**.** *A drift diffusion process with drift μ and variance parameter σ^2 is a family of random variables $(X_i)_{i\in\mathbb{R}^+}$ which satisfies the conditions from the previous definition except for the third which is replaced by:*

(3′) $Y_t - Y_s \sim \mathcal{N}((t-s)\mu, (t-s)\sigma^2)$*.*

Using a stochastic differential equation such a process can also be written as follows, where W denotes a Wiener process.

$$dX = \mu\, dt + \sigma\, dW \tag{2:37}$$

2.5.2.3 *Relation to hypothesis testing and application to experimental data*

In the present section, drift diffusion processes are related to hypothesis testing problems. Furthermore, it is explained briefly how such processes can be used as a model of human decision making, specifically in two alternative forced choice (2AFC) tasks.

A decision procedure that is equivalent to the SPRT is obtained by taking the logarithm of equation (2:32), such that the *log likelihood ratio* (LLR) after the *i-th* observation is:

$$LLR_i = \sum_{1 \leq j \leq i} \log L(y_j).$$

More observations are sampled as long as $\log \underline{\tau} \leq LLR_i \leq \log \overline{\tau}$. As soon as one of the decision boundaries, $\log \underline{\tau}$ or $\log \overline{\tau}$, is crossed, a decision is made according to a decision rule that is completely analogue to 2:33. Thus, we can conceive of the SPRT as the random walk:

$$\begin{aligned} Z_0 &= 0, \\ Z_i &= Z_0 + \sum_{1 \leq j \leq i} \log L(Y_j), \qquad \text{for } i \geq 1. \end{aligned} \tag{2:39}$$

Moreover, if the i.i.d. random variables $\log L(Y_j)$ have mean μ and variance σ^2, then the random walk in 2:39 converges to its *continuum limit* X_t as defined by the equation in 2:37, after a change of scale (see Bogacz et al., 2006, Appendix A). Therefore, we can also think of the optimal decision procedure for sequential hypothesis testing problems as a drift diffusion process.

Stochastic processes have been used to model human descion making at least since Stone (1960). In his *drift diffusion model* (DDM), Ratcliff (1978) famously proposed to model actual decision processes performed by humans in psychological experiments using drift diffusion processes. The application of the drift diffusion model (DDM) and similar models in cognitive science was recently reviewed by Forstmann, Ratcliff, and Wagenmakers (2016). Given a set of free model parameters, the DDM allows us to predict proportions of errors, the expected *decision time* and the distribution of *decision times*. For example, if we assume no response bias, which is appropriate in

a situation with equal priors and uniform cost assignment, the proportion of errors (denoted ER) is predicted to be:

$$ER = \frac{1}{1 + \exp\left(\mu z / \sigma^2\right)} \tag{2:40}$$

and the mean decision time (DT) is predicted to be

$$DT = \frac{z}{\mu} \tanh\left(\frac{\mu z}{\sigma^2}\right), \tag{2:41}$$

where $z = \frac{1}{2}(\overline{\tau} - \underline{\tau})$ is the distance from the so-called starting point, i.e. 0, to the (symmetric) decision boundaries (for details see e.g. Bogacz et al., 2006).

More precisely, this model is appropriate in the so-called free response paradigm, in which the participants of an experiment determine the time when the decision is made. If the decision time is controlled by the experimenter, e.g. via a response signal, the decision process can be modeled analogously to the LRT from section 2.5.1. In that case, ER is predicted as (again, see Bogacz et al., 2006, for details):

$$ER = \Phi\left(-\frac{A}{\sigma}\sqrt{DT}\right), \quad \text{where } \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt. \tag{2:42}$$

2.6 GRADABLE ADJECTIVES AND COMPARATIVES

In chapters 5 and 6, compartive quantifiers like *more than five* or *fewer than half of the* are analyzed using the syntax and semantics of ordinary comparatives, as was suggested by Hackl (2000). Ordinary comparatives involve *gradable adjectives* like *tall*, *heavy* or *old*. The ability of adjectives like these to appear in comparatives is even considered a diagnostic for their gradability (Lassiter, 2015). Another robust one is their ability to combine with degree modifiers, e.g. *very tall*, *somewhat heavy* or *quite old*. Some basic aspects of the syntax and semantics of gradable adjectives and comparatives are briefly summarized in the present section.

There are two prominent approaches to the semantics of gradable adjectives. The first is called the *vague predicate analysis*. It is based on the idea that gradable adjectives denote special kinds of predi-

cates, namely vague ones, that are interpreted via partial, context-dependent truth assignments. These truth assignments are subject to certain consistency constraints and thus induce an ordering over their domain (Klein, 1980 who refers to a paper by Hans Kamp). The second approach is referred to as the *scalar analysis*. It is based on mappings from individuals to *degrees*, i.e. elements of an ordered set (Cresswell, 1976). Here, we focus on the latter type of approach. For comparison of the two and discussion about how they are related the reader is referred to von Stechow (1984), Kennedy (1997) and Lassiter (2015).

2.6.1 *Degrees*

Scalar analyses of gradable adjectives are based on a special kind of objects called *degrees*. The intuition is that degrees are certain measurements of individuals (e.g. height, weight, age, *etc.*). More formally, they are elements of a *scale* which is defined as a triple $\langle D, \geq, \text{DIM} \rangle$, where D is a set, $\geq$ is a reflexive, antisymmetric and transitive relation over D and DIM is the dimension of D (e.g. height in meters, weight in kg, age in years, *etc.*). The semantic type of degrees is denoted by $\mathtt{d}$. There is much more to say about the type of scales that are relevant to the semantics of natural language (For recent discussions see Kennedy & McNally, 2005a; Fox & Hackl, 2007 and Solt, 2015), but we will leave it at this brief introduction. Moreover, we will usually ignore dimensions, assume that degrees are totally ordered and assume that they are equipped with an addition operation denoted by $+$. Specifically, we think of a scale as a *totally ordered group* $(D, +, \geq)$. Going even further, we can often simply think of scales as structures that are isomorphic to the real numbers.

2.6.2 *Two types of scalar analyses*

In 2:43-a the scalar adjective *tall* appears in what is called its *positive* form. Scalar analyses assign an interpretation as indicated in the paraphrase, 2:43-b, to sentences like this. Comparative sentences like 2:44-a are analyzed as exemplified in 2:44-b.

(2:43) a. Simon is tall.

b. Paraphrase: The degree to which Simon is tall exceeds the range of 'neutral' degrees of tallness in the relevant comparison class.

(2:44) a. Elin is taller than Simon (is).
b. Paraphrase: The degree to which Elin is tall exceeds that to which Simon is tall.

There are two prominent variants of the degree-based approach to gradable adjectives and comparatives. The first takes gradable adjectives to denote binary relations between individual and degrees (Cresswell, 1976,von Stechow, 1984, see also Beck, 2011). The second approach assumes gradable adjectives to denote functions that map individuals to degrees, called *measure functions* (Kennedy, 1997 and reference therein, specifically Bartsch and Vennemann). These two approaches result in slightly different lexical entries, syntactic structures and semantic mechanisms. The question which approach, if any of the two, is empirically adequate is not settled yet. One crucial difference between the two approaches lies in what predictions follow regarding the possibility of certain scope ambiguities (for discussion see Kennedy, 1997 and I. Heim, 2000) and the syntactic and semantic status of *than*-phrases (Bhatt & Pancheva, 2004; Grosu & Horvath, 2006). Both approaches are illustrated briefly here by means of the examples in 2:44 and 2:43.

2.6.2.1 *The relational approach*

The relational approach conceives of gradable adjectives like *tall* as binary relations between degrees and individuals (see first row of Table IVa). As the following example shows, the degree argument may be overtly realized in form of a *measure phrase* like *85 cm*. In that case, it combines with the adjective to form an AdjP.

(2:45) Simon is 85 cm tall.

THE POSITIVE. In sentences like 2:44 there is no overt degree argument. It is commonly assumed that a phonologically silent degree quantifier called POS takes its place (see Figure Va). This quantifier provides an implicit *standard of comparison*. The first row of Table IVa shows a lexical entry of POS. The predicate $\mathbf{ntr}_{Q,C}$ is meant to encode a 'neutral interval' of degrees with respect to a *degree predi-*

Table IV

Some lexical entries.

(a) *relational approach*

tall	ADJ	`<d,<e,t>>`	$\lambda d.\lambda x.\mathbf{height}(x) \geq d$
POS	DEG	`<<d,t>,t>`	$\lambda Q.\forall d'\left(\mathbf{ntr}_{Q,C}\ d' \rightarrow Q\ d'\right)$
-er	DEG	`<<d,t>,<<d,t>,t>>`	$\lambda P.\lambda Q.\exists d'\left(\neg P\ d' \wedge Q\ d'\right)$

(b) *functional approach*

tall	ADJ	`<d,e>`	**height**
POS	DEG	`<<e,d>,<d,t>>`	$\lambda f.\lambda x.f(x) > \mathbf{s}(f,c)\big)$
-er	DEG	`<<e,d>,<<<e,d>,d>,<e,t>>>`	$\lambda f.\lambda G.\lambda x.f(x) > G(f)$

cate Q (type `<d,t>`) and a contextually determined *comparison class* C (type `<e,t>`; for technical details and discussion see e.g. von Stechow, 1984; Kennedy, 2007). In order to achieve an interpretable logical form (LF), POS is assumed to undergo QR (see I. Heim & Kratzer, 1998). In order for sentence 2:44-a to be true, Simon's height has to be an upper bound of the neutral interval. Figure V illustrates the syntax and semantics of sentence 2:44-a according to the relational approach. For more details the reader is referred to the cited literature. Of course, some relational proposals in the literature differ slightly from what is presented here. For example, one common alternatives for the lexical semantics of POS is something along the lines of: $\lambda Q.\textsc{max}(\lambda d.Q\ d\big) > \mathbf{s}$, where $\mathbf{s}$ is the contextually determined standard of comparison. The version chosen here is a notational variant of the subset semantics: $\lambda Q.N_{Q,C} \subseteq Q$ (cf. von Stechow, 1984; I. Heim, 2006; Solt, 2009; Beck, 2011), which is more general than the version using MAX.

THE COMPARATIVE. The comparative morpheme *-er* relates two degree predicates. A lexical entry is provided in the third row of Table IVa. [4] At LF, *-er* composes with the *than*-phrase, which yields a degree quantifier of type `<<d,t>,t>`. This expression undergoes QR – just like POS – to obtain an interpretable structure. Figure VI shows the syntactic structure and translation to a typed language. The surface word order is derived from the syntactic structure in VIa by way of two further steps. Firstly, the *than*-phrase is extraposed.

[4] Note that according to this analysis *-er* corresponds to a non-conservative GQ (cf. Bhatt & Pancheva, 2004; Grosu & Horvath, 2006), which is surprising given the fact that such GQs are not found in the nominal domain.

Secondly, a morphological rule has to apply that turns *-er tall* into *taller*.

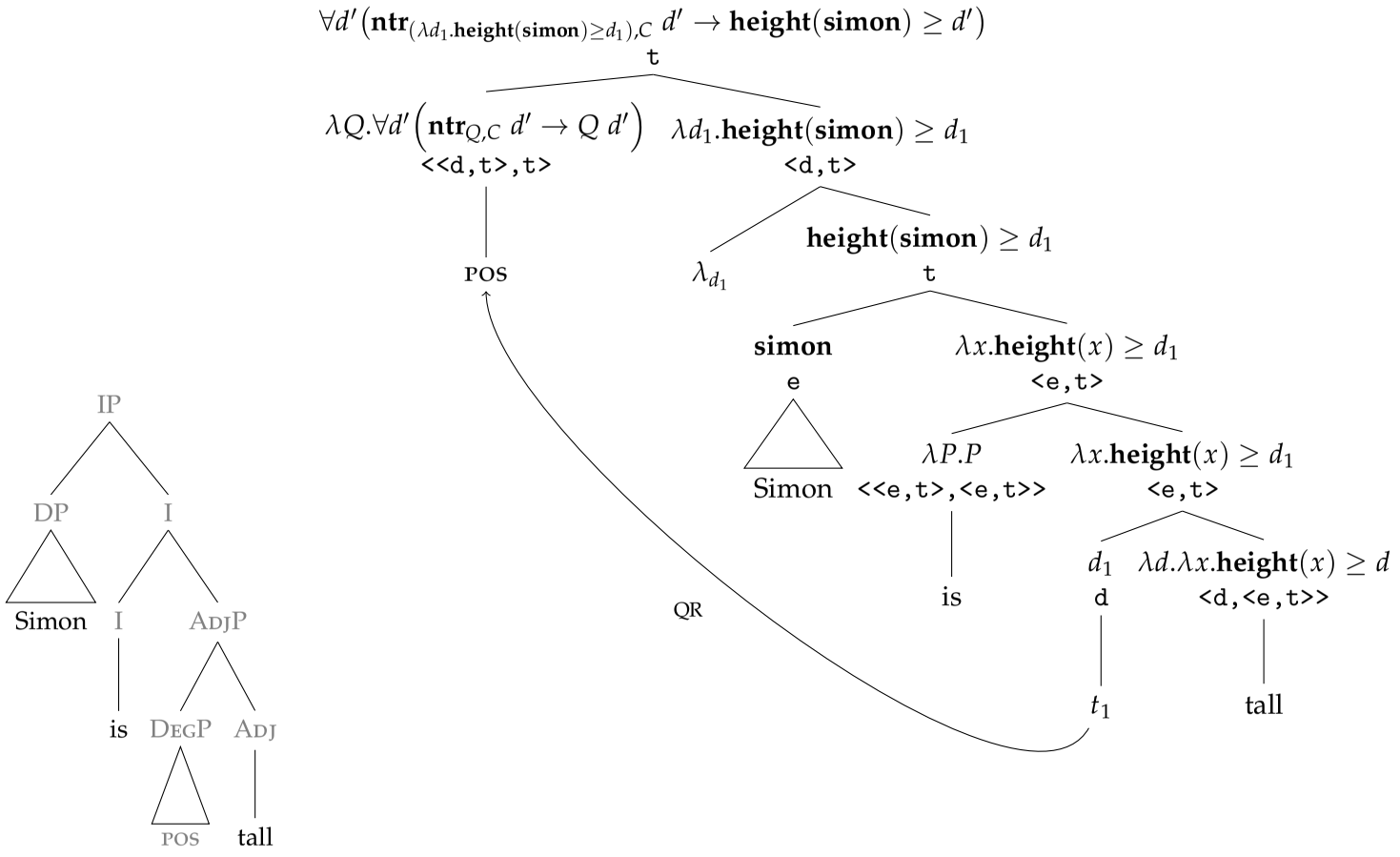
2.6.2.2 *The functional approach*

The functional approach identifies gradable adjectives with *measure functions* that map individuals to degrees. This is shown in the first row of Table IVb. To obtain a predicate and eventually a proposition from this denotation, degree morphology, i.e. lexical material of category DEG, like POS or *-er*, is crucial. A syntactic difference between the functional and the relational approach is that, in the former, the degree expression is considered the head and the adjective the complement whereas the latter assumes opposite roles.

THE POSITIVE. In the positive form, the value of a measure function f, e.g. Simon's height, is compared to the standard $\mathbf{s}(f, c)$. This is straightforwardly achieved using the lexical entry in the second row of Table IVb. As above, the standard depends on the compositionally provided measure function and a contextually determined comparison class c. In contrast to the relational approach, no QR takes place. Detailed derivations are shown in Figure VII.

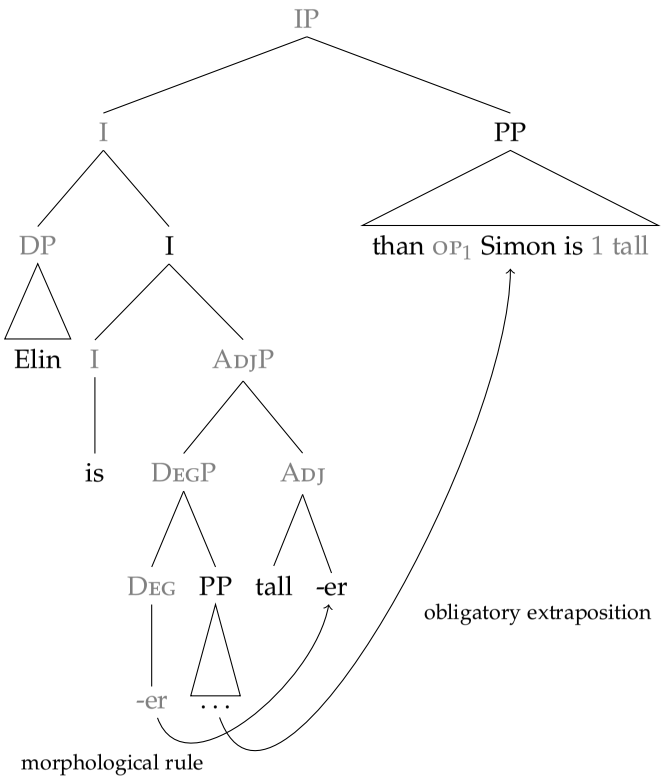
THE COMPARATIVE. A lexical entry of the comparative morpheme is given at the bottom of Table IVb. As shown in Figure VIII comparatives are interpreted *in situ*, without QR. The interpretation of the PP is simplified in the figure. Originally, Kennedy (1997) proposed that the PP is interpreted as $\lambda g.\textsc{max}\big(\lambda d.g(\mathbf{simon}) \geq d\big)$. For our example sentence, the end result is, however, the same. Another point I want to highlight is that the denotation of *-er* is tailored to the case of *comparative deletion*. In the case of *comparative subdeletion*, as in *the table is wider than the door is high*, for example, Kennedy proposes a different denotation of *-er*: $\lambda f.\lambda d.\lambda x.f(x) > d$.[5]

[5] He stressed that these variants make use of different combinatorics but do not lead to any truth-conditionally relevant differences. I would like to add that the denotation in Table IVa can, in fact, be derived from the just-mentioned one by application of the type shifter $\lambda F.\lambda r.\lambda s.F(r)(s(r))$, which is discussed by Jacobson and Barker (2005) under the name **s**.

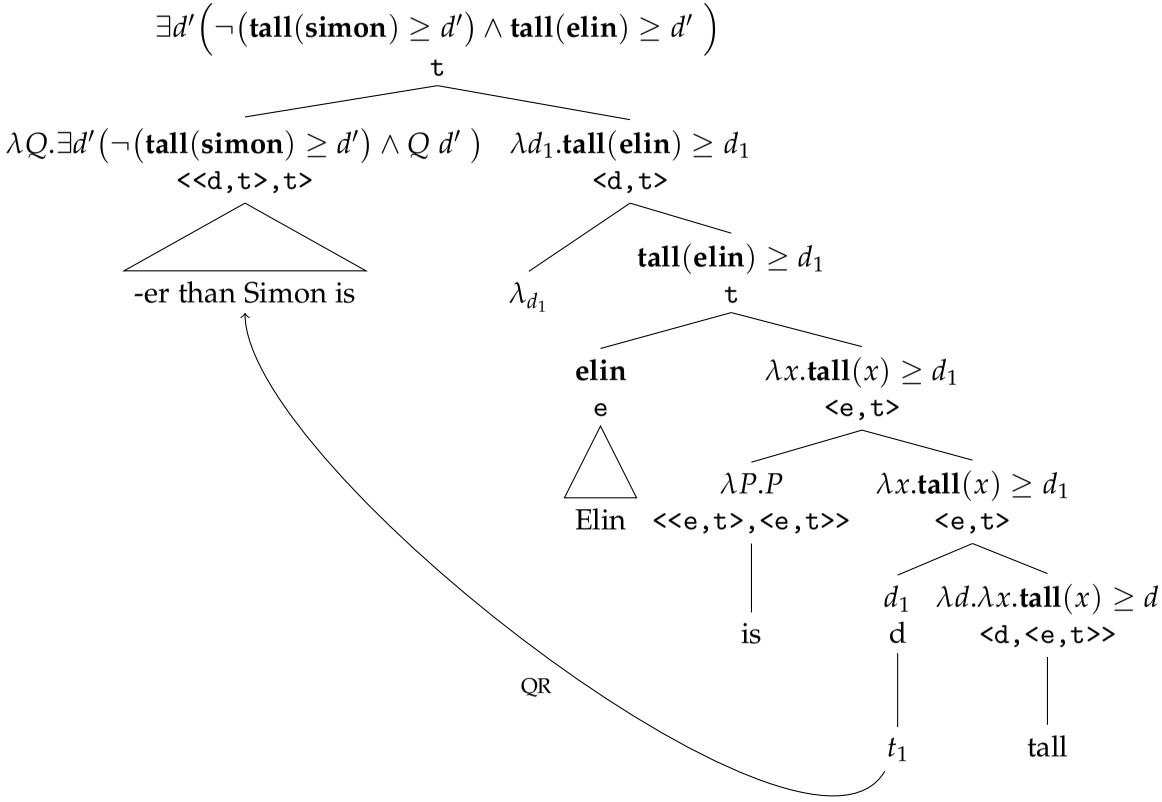


(a) Syntactic structure *(b)* Translation into a typed language

Figure V. The positive under a relational approach

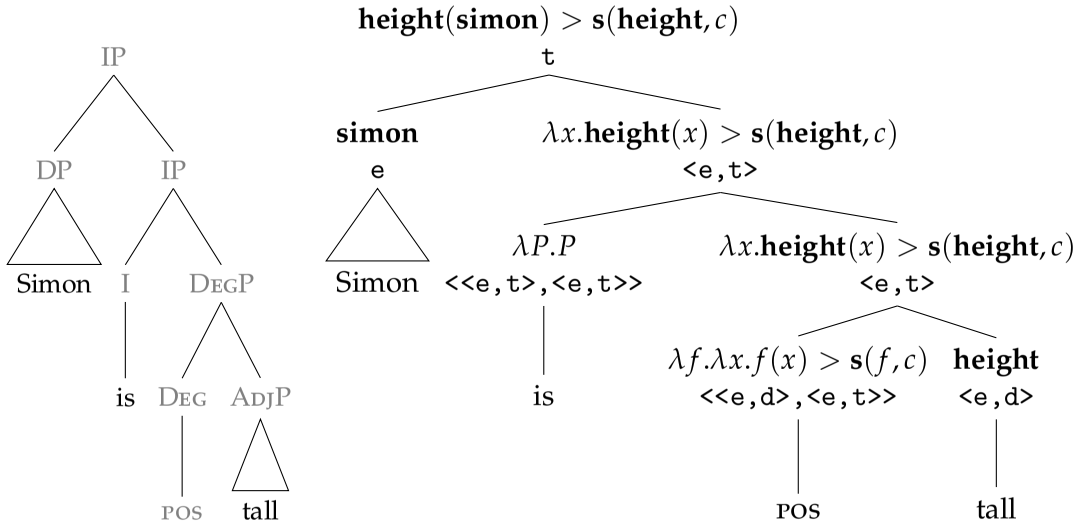


(a) Syntactic structure



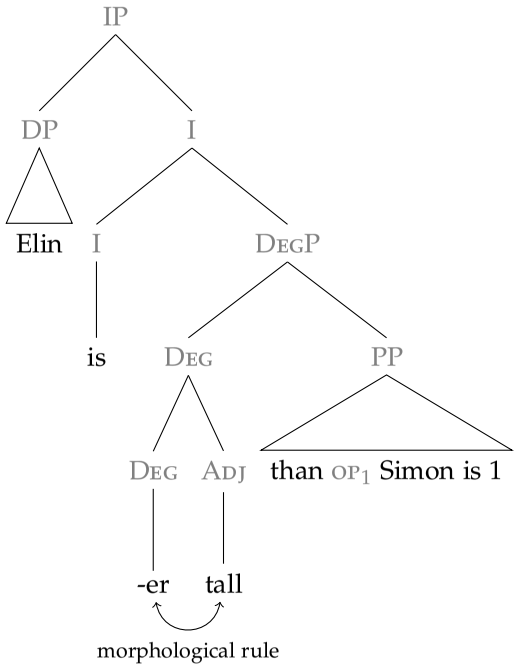
(b) Translation to a typed language

Figure VI. The comparative under a relational approach.

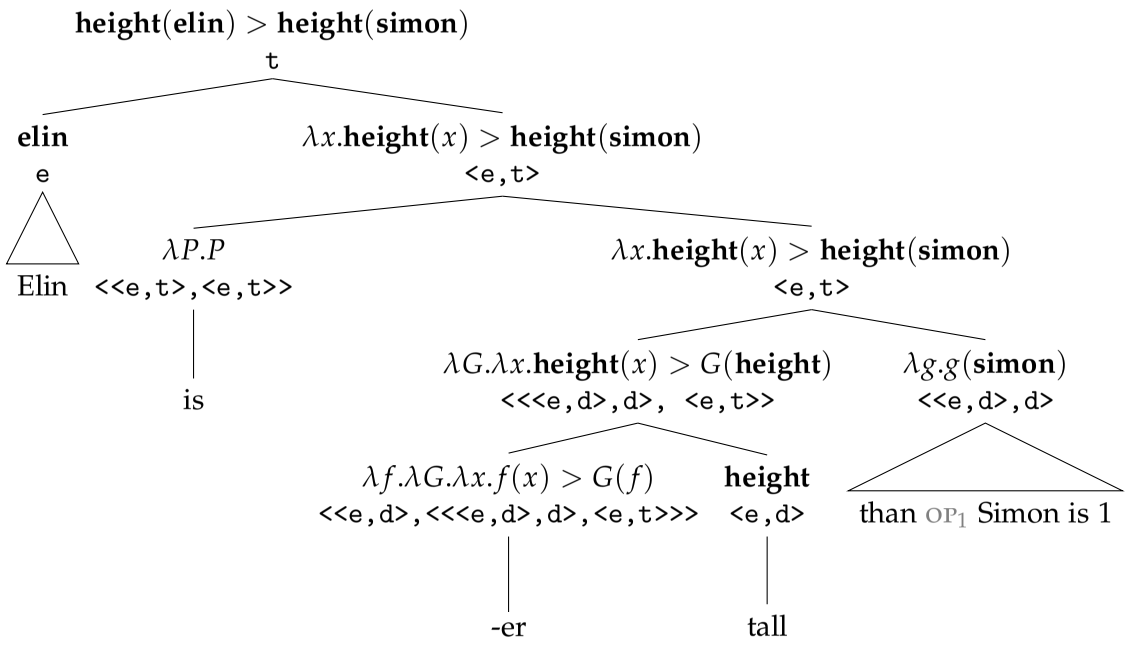


(a) Syntax structure *(b)* Translation into a typed language

Figure VII. The positive under a functional approach.



(a) Syntax structure



(b) Translation to a typed language

Figure VIII. The comparative under a functional approach.

2.7 COMBINATORY CATEGORIAL GRAMMAR

In chapters 5 and 6, *combinatory categorial grammar* (CCG; Ades & Steedman, 1982 and Szabolcsi, 1987) is used to derive symbolic meaning representations of comparative quantifiers. The basics of this formalisms are introduced here.

CCG is a grammar formalism in which syntax and semantics go hand in hand (for an introduction, see e.g. Steedman & Baldridge, 2011). It derives syntactic well-formedness, and implicitly also syntactic structure, in tandem with symbolic meaning representations. Furthermore, it is a lexicalist formalism that assumes only a small set of derivation rules and, apart from these, encodes linguistic regularity in the lexicon. Moreover, the formalism is relatively expressive but still allows for efficient parsing (see Steedman, 2000, part III or Steedman & Baldridge, 2011 for discussion). From a psycholinguistic perspective, CCG is especially interesting because, in many cases, it allows for strictly incremental, i.e. word-by-word, syntactic parsing and semantic interpretation (for discussion see e.g. Altmann & Steedman, 1988; Steedman, 2000; Demberg, 2012).

As in many other formalisms, lexical items have a syntactic category and a semantic denotation in CCG. Syntactic categories may be primitive (e.g. S or NP) or complex (e.g. S/NP). Complex categories encode potential combinations with other categories. For instance, the category S/NP is looking to combine with an expression of category NP to its right to yield a sentence, i.e. an expression of category S. The category S\NP, where the slash points in the opposite direction, is looking to combine with an NP to its left to form a sentence. On the semantic side, expressions with category S/NP or S\NP denote *functors* that apply to *arguments* denoted by expressions of category NP. With the exception of a few unary type-shifting rules, derivation rules combine two adjacent expressions to form an expression with new syntactic category and semantic denotation. *Combinators* (i.e. lambda expression with no free variables) are used to combine denotations. One example is the combinator **B**, which is defined by $\mathbf{B}fga := f(ga)$ and thus corresponds to the lambda-term $\lambda x.\lambda y.\lambda z.x(yz)$. The combinator **B** is used in so-called function composition rules. The rules and combinators that are standardly used in CCG are given below. In general, left-associative notation is used in all rules and derivations, i.e. $(xy)z$ is written as xyz. In the rules

below, The symbols X, Y and Z denote variables over syntactic categories and f, g, a range over denotations.

There are two rules for functional application. In the first, the functor precedes its argument. In the second, the argument precedes the functor. They correspond to the above examples.

(2:46) *Functional application rules*

a. *Forward functional application*
$$X/Y : f \quad Y : a \quad \Rightarrow_{>} \quad X : fa$$
b. *Backward functional application*
$$Y : a \quad X\backslash Y : f \quad \Rightarrow_{<} \quad X : fa$$

The second type of rules implements function composition. It is based on the combinator **B**. There are four basic composition rules, which are shown in 2:47. Moreover, these four rules can be generalized as in 2:48.

(2:47) *Composition rules*

a. *Forward harmonic composition*
$$X/Y : f \quad Y/Z : g \quad \Rightarrow_{>\mathbf{B}} \quad X/Z : \lambda x.f(gx)$$
b. *Backward harmonic composition*
$$Y\backslash Z : g \quad X\backslash Y : f \quad \Rightarrow_{<\mathbf{B}} \quad X\backslash Z : \lambda x.f(gx)$$
c. *Forward crossing composition*
$$X/Y : f \quad Y\backslash Z : g \quad \Rightarrow_{>\mathbf{B}_\times} \quad X\backslash Z : \lambda x.f(gx)$$
d. *Backward crossing composition*
$$Y/Z : g \quad X\backslash Y : f \quad \Rightarrow_{<\mathbf{B}_\times} \quad X/Z : \lambda x.f(gx)$$

(2:48) *Generalized composition rules*

a. *Generalized forward harmonic composition*
$$X/Y : f \quad Y/Z_1/Z_2 : g \quad \Rightarrow_{>\mathbf{B}^2} \quad X/Z_1/Z_2 : \lambda x.\lambda y.f(gxy)$$
b. *Generalized backward harmonic composition*
$$Y\backslash Z_1\backslash Z_2 : g \quad X\backslash Y : f \quad \Rightarrow_{<\mathbf{B}^2} \quad X\backslash Z_1\backslash Z_2 : \lambda x.\lambda y.f(gxy)$$
c. *Generalized forward crossing composition*
$$X/Y : f \quad Y\backslash Z_1\backslash Z_2 : g \quad \Rightarrow_{>\mathbf{B}^2_\times} \quad X\backslash Z_1\backslash Z_2 : \lambda x.\lambda y.f(gxy)$$
d. *Generalized backward crossing composition*
$$Y/Z_1/Z_2 : g \quad X\backslash Y : f \quad \Rightarrow_{<\mathbf{B}^2_\times} \quad X/Z_1/Z_2 : \lambda x.\lambda y.f(gxy)$$

Finally, there is a forward and a backward type-raising rule. These two rules make use of the combinator **T**, which is defined by $\mathbf{T}af := fa$ and corresponds to the lambda term $\lambda x.\lambda y.yx$

(2:49) *Type-raising rules*

a. *Forward type-raising*

$$X : g \quad \Rightarrow_{>\mathbf{T}} \quad Y/(Y\backslash X) : \lambda f.fg$$

b. *Backward type-raising*

$$Y : g \quad \Rightarrow_{<\mathbf{T}} \quad Y\backslash(Y/X) : \lambda f.fg$$

In Figure IX, an example CCG derivation is shown. The first line of the derivation corresponds to lexical retrieval. Rule applications are written horizontally instead of vertically. For example, application of the rule in 2:49-a is written as:

$$\dfrac{\begin{array}{c}X\\g\end{array}}{\begin{array}{c}Y/(Y\backslash X)\\ \lambda f.f(g)\end{array}}{>}\mathbf{T}$$

Below, in chapters 5 and 6, a special type of lexical retrieval is used. In particular, lexical items are decomposed into several semantic building blocks that can independently interact with their local environment. Moreover, rudimentary use of *features* is made to rule out ungrammatical derivations. Specifically, features will be used to encode that the comparative morpheme selects for a *than*-phrase. A typical example of the usage of features is case marking. For example, the category NP_{acc} carries the *accusative* feature and $\mathrm{S}\backslash\mathrm{NP}_{nom}/\mathrm{NP}_{acc}$ would be the category of a transitive verb that selects for a nominative subject and an accusative object.

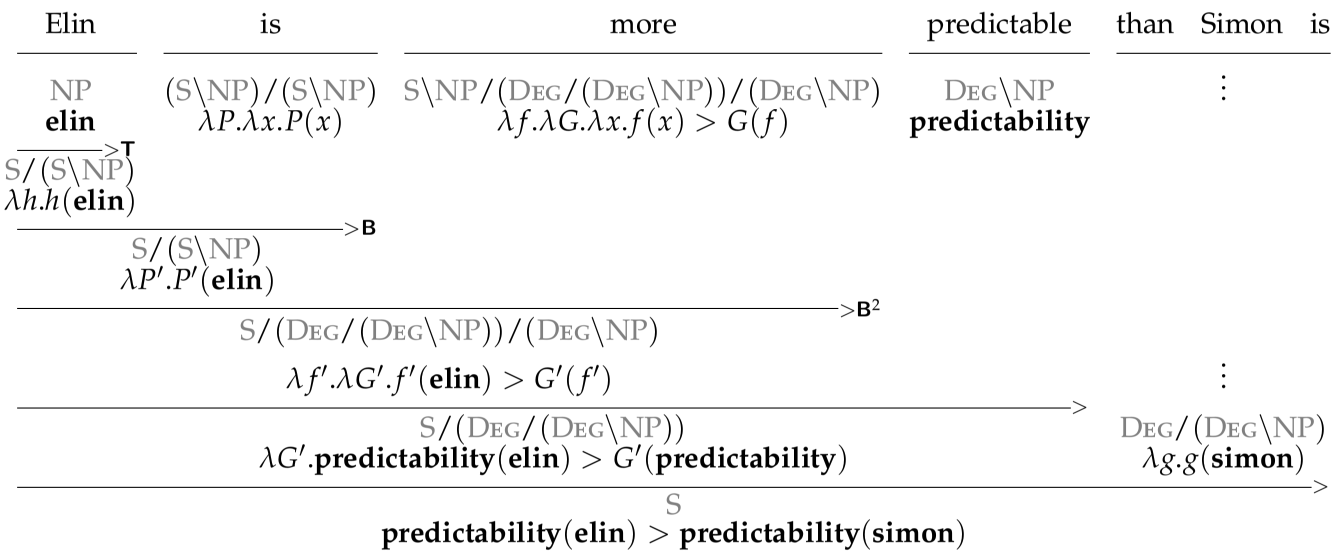


Figure IX. Partially incremental CCG derivation based on the account of Kennedy (1997).

3

EASY SOLUTION TO A HARD PROBLEM? COMPUTATIONAL COMPLEXITY AND PROCESSING DIFFICULTY OF QUANTIFIED RECIPROCALS[1]

A first step towards cognitively plausible semantic theories would be to establish a link between the truth conditions of a sentence and how it is processed. Approaching this problem from an abstract point of view, we may claim that a theory of cognition can only be realistic if computations are bounded in computational complexity since they have to be performed in real time by agents with limited processing resources. Considerations about computational complexity have guided theoretical work on a number of phenomena in cognitive science, including, for instance, analogical reasoning (e.g. Veale & Keane, 1997), Bayesian inference (e.g. Cooper, 1990), motion planning (e.g. Joseph & Plantings, 1985 or Reif, 1985), intentional communication (e.g. van Rooij et al., 2011) and parsing (e.g. Barton, Berwick, & Ristad, 1987, Ristad, 1993 or Wareham, 1999). In line with this type of considerations, Frixione proposed the following heuristic for cognitive science:

> [T]o build a theory of competence for a certain cognitive task, a computationally tractable function modelling such a task should be singled out. (Frixione, 2001, p.17)

The term *computationally tractable function* is used here to refer to functions that can be computed in *polynomial time* by a deterministic Turing machine (see Definition 2:17), a class of functions called **P** (see Arora & Barak, 2009 and also section 2.4, Definition 2:21). Iris van Rooij (2008) analyzed the proposal of Frixione and similar work as embracing a two-part hypothesis. Firstly, there is the informal *tractable cognition hypothesis* which states the following.

[1] This chapter is a modified version of Schlotterbeck and Bott (2013). Parts of it were also presented at the "Workshop on Logic and Cognition" at ESSLLI 2013, in Opole.

Hypothesis 3:1 (Tractable cognition hypothesis, van Rooij, 2008)**.** *Human cognitive capacities are constrained by the fact that humans are finite systems with limited resources for computation.*

Secondly, the more specific *P-cognition thesis* (PCT) explicitly restricts cognitive capacities to a precisely defined class of functions, namely the just mentioned **P**.

Hypothesis 3:2 (P-cognition thesis, PCT, van Rooij, 2008)**.** *Human cognitive capacities are polynomial-time computable.*

Contrary to other areas of cognitive science – also including other areas of linguistics – computational complexity has only recently begun to receive attention in formal semantics. Notable examples include the work of Mostowski (1998), Mostowski and Szymanik (2007, 2012), and Szymanik (2009, 2010, 2016). In line with the classical idea that the meaning of a sentence is characterized by its truth conditions, these authors identify the meaning of a natural language sentence with the class of its models, i.e. situations in which the sentence is true. They investigated the computational complexity of the decision problem whether a model belongs to this class or not.

What predictions for semantic processing follow from these considerations? It is clear that, according to the strictest reading of the PCT, deciding about the truth or falsity of computationally intractable truth conditions should not be a possible cognitive function. [2] Alternatively, less strict variants of the PCT are conceivable as well (see van Rooij, 2008, for an approach based on *fixed parameter tractability*; see also section 3.3). These would basically predict that intractable truth conditions can be evaluated as long as certain parameters of the models stay within certain bounds.

In order to test these predictions, the present chapter investigates the semantic interpretation of ambiguous sentences with one intractable reading and tractable alternatives. In particular, the PCT is applied to semantic processing by looking at a particularly interesting test case that was originally discussed by Szymanik (2009). He considered the interpretation of reciprocal sentences with generalized quantifiers as antecedents. These sentences are of the form DET N V *each other*, or a closely related one. What makes these sentences particularly interesting for the present purposes is the following. For some quantificational DETS (e.g. the proportional *most*), the evaluation of

[2] Here, it is tacitly assumed that $\mathbf{P} \neq \mathbf{NP}$. Problems in **P** are called tractable. Problems that are **NP**-hard are called intractable.

one of their readings is **NP**-complete (see Definition 2:23) whereas for others (e.g. *all*) all readings are known to be in **P**. Our experiments tested whether the possible readings are limited to those which are known to be in **P** (i.e. computationally tractable) as proposed by Szymanik. In particular, it was investigated whether comprehenders shift from a reading that is preferred under normal circumstances to an otherwise dispreferred but computationally tractable one in order to avoid having to deal with an intractable verification problem.

Beside the evaluation of truth conditions in a model, cognitively plausible theories about semantic processing should encompass other aspects of interpretation, too. These include the recognition of logical relations – such as entailments or contradictions – between sentences and also comprehension, which includes, for example, the derivation of a sentence's the truth conditions from the meanings of its parts. With regard to computational complexity, it is an open theoretical and empirical question how these different aspects are related (see e.g. Kirousis & Kolaitis, 2001; Pratt-Hartmann, 2010, for related theoretical investigations). Here, we focus on possible connections between verification and comprehension. We assume that deriving the truth conditions of a sentence is a prerequisite for its verification. Therefore, comprehension has to be accounted for when we aim to investigate the evaluation of truth conditions in detail.

In principle, comprehension could be unaffected by the computational tractability of the corresponding verification problem. This would, however, lead to situations in which a comprehender may face an intractable problem when relating the sentence to a concrete state of affairs. It is plausible that the processing system is set up in a way to avoid this kind of situation whenever possible. There are at least two related avoidance strategies.

Hypothesis 3:3 (Avoidance Strategy A)**.** *The processor keeps track of particular instances of intractable sentences and, once they can be related to a new instance of the same problem, tries to shift the meaning to a simpler interpretation. Specifically, we may expect to see these shifts if a comprehender encounters an ambiguity in which a tractable meaning exists besides an intractable alternative.*

Hypothesis 3:4 (Avoidance Strategy B)**.** *The internal language in which sentence meaning is represented is restricted in such a way that intractable meanings are excluded automatically.*

In the following, investigations of both, verification and comprehension are reported. Verification was studied in a sentence-picture verification task experiment where participants had to decide whether a sentence matches a graphically depicted model or not. Comprehension was studied in a picture completion experiment in which a model had to be drawn for a given sentence.

The remaining part of the chapter is structured as follows. The next section provides the necessary theoretical background and discusses related empirical work. Section 3.4 presents a pretest establishing that the materials to be tested in the main experiments fulfill certain semantic requirements. The picture completion experiment which was conducted to elicit the preferred interpretation of quantified reciprocal sentences with or without an intractable reading is reported in section 3.5. Section 3.6 presents the picture verification experiment that tested effects of computational complexity during verification. In section 3.7, we conclude with a discussion of whether the findings are compatible with the PCT and related hypotheses.

3.1 QUANTIFIED RECIPROCALS AND THE STRONGEST MEANING HYPOTHESIS

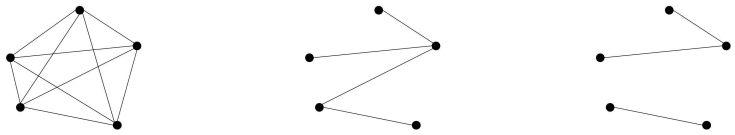
The sentence in 3:5 exemplifies quantified reciprocals. Glossing over syntactic details, they follow the schema DET *A R each other*, where DET is a quantificational determiner, *A* denotes a unary relation (e.g. a noun) and *R* stands for a binary relation (e.g. a transitive verb). In constructions of this kind, we call DET the *quantificational antecedent*, *A* the *restriction* and *R* the *reciprocal relation*.

(3:5) All dots are connected to each other.

As was analyzed thoroughly by Dalrymple, Kanazawa, Kim, McHombo, and Peters (1998), a sentence like 3:6 may receive one of several readings. Some of these are captured by the following logical formulas. In 3:6-b, the formula $\phi[x, y]$ is meant to expresses that x is reachable from y, i.e. there is a path between them (see Ajtai & Fagin, 1990, for the details of this lengthy formula).

(3:6) a. $\forall x\Big(\mathbf{dot}\ x \rightarrow \big(\forall y(\mathbf{dot}\ y \rightarrow (x \neq y \rightarrow \mathbf{connected}\ xy))\big)\Big)$

b. $\forall x\Big(\mathbf{dot}\ x \rightarrow \big(\forall y(\mathbf{dot}\ y \rightarrow (x \neq y \rightarrow \phi[x, y]))\big)\Big)$



(a) graph reading *(b)* path reading *(c)* pair reading

Figure X. Possible models for the quantified reciprocal in example 3:5. The vertices in the graphs represent elements of the restriction, A. The edges represent the reciprocal relation, R. The graph in a is a model of the formula in 3:6-a, b is a model of 3:6-b and c of 3:6-c.

c. $\forall x\Big(\mathbf{dot}\ x \rightarrow \Big(\exists y\big(\mathbf{dot}\ y \rightarrow (x \neq y \rightarrow \mathbf{connected}\ xy)\big)\Big)\Big)$

The formula in 3:6-a states that any two members of the restriction (*dots*) participate in the reciprocal relation (*to be connected to*). By contrast, 3:6-b states that any two members of the restriction are connected by a path. In the present example, such a path would amount to an indirect connection. Quantified reciprocals like the example 3:5 may also exhibit a third reading given in 3:6-c, which states that for any member of the restriction there is some other member and both participate in the reciprocal relation. We refer to the first reading, 3:6-a, as *complete graph reading*, the second, 3:6-b, as *path reading* and the third, 3:6-c, as *pair reading*. Figure X shows possible models for each of the formulas in 3:6.

As noted by Dalrymple et al., reciprocal expressions like *each other* can be analyzed as generalized quantifiers (GQs) of type $(1,2)$ (see section 2.2, Definition 2:1). Moreover, concerning quantified reciprocals, the truth conditions in 3:6 can be considered a special case of the second order formulas in 3:7 (adopted from Szymanik, 2009 with slight modification). If we assume that in these formulas Q stands for the type $(1,1)$ quantifier corresponding to *all*, then 3:7-a–c would be equivalent to 3:6-a–c, respectively.

(3:7) a. $\exists X \subseteq A\Big(\big(\mathrm{Q}v\ Av, Xv\big) \wedge \big(\forall x, y \in X\,(x \neq y \rightarrow Rxy)\big)\Big)$

b. $\exists X \subseteq A\Big(\big(\mathrm{Q}v\ Av, Xv\big) \wedge \big(\forall x, y \in X\,(x \neq y \rightarrow \phi[x,y])\big)\Big)$

c. $\exists X \subseteq A\Big(\big(\mathrm{Q}v\ Av, Xv\big) \wedge \big(\forall x \in X\ \exists y \in X\,(x \neq y \wedge Rxy)\big)\Big)$

Each of these formulas defines a GQ of type $(1,2)$ obtained from Q via an operation called *Ramseyfication* (Hella, Väänänen, & West-

erståhl, 1997; Szymanik, 2009). They are equivalent to what Szymanik (2009, p. 238) called

$$\begin{aligned}&\big(Ram_S(Q)\big)_M(A,R),\\&\big(Ram_I(Q)\big)_M(A,R) \text{ and}\\&\big(Ram_W(Q)\big)_M(A,R), \text{ respectively,}\end{aligned}$$

where $Ram_S(Q)$, $Ram_I(Q)$ and $Ram_W(Q)$ are the mentioned type $(1,2)$ quantifiers. The subscript S stands for *strong*; I stands for *intermediate* and W stands for *weak*.

There are various approaches to reciprocals in general and quantified reciprocals specifically (among others: Langendoen, 1978; I. Heim, Lasnik, & May, 1991; Sternefeld, 1998; Beck, 2001; Sabato & Winter, 2012). An advantage of some other approaches is that they derive truth conditions compositionally, e.g. by means of logical machinery that is independently motivated by the syntax and semantics of plurals. Compositionality is, however, irrelevant for our present purpose because we are mainly interested in the truth conditions of complete sentences and how difficult they are to verify. There seems to be general agreement that 3:8-a–c are possible readings of quantified reciprocals.

The labels on Ram_S, Ram_I, Ram_W already indicate the logical dependencies between these readings. In fact, if Q is upward entailing (UE, see Definition 2:8), then the complete graph reading is the logically strongest interpretation which implies all the others. [3] To account for interpretation preferences of sentences like these, Dalrymple et al. have put forward the *strongest meaning hypothesis (SMH)*.

Hypothesis 3:9 (Strongest meaning hypothesis, SMH, Dalrymple et al., 1998)**.** *Quantified reciprocals receive the logically strongest interpretation consistent with the reciprocal relation and relevant background information.*

For example, the SMH captures the intuition that the complete graph reading is the preferred reading of the example 3:5, above. Moreover, it also captures the difference in interpretation between the following two sentences.

[3] More precisely, this is the case if there are at least two elements in the quantificational restriction. Dalrymple et al. stated this condition explicitly in their truth conditions. With UE quantifiers, this is, however, often irrelevant, because the quantificational antecedent requires a restriction of at least two elements – either because of its semantics or on pragmatic grounds.

(3:10) a. All the students knew each other.
b. All the students followed each other into the room.

The preferred interpretation of 3:10-a is the complete graph reading, whereas 3:10-b receives a path interpretation. The SMH is not undisputed, however. Kerem, Friedmann, and Winter (2011) have presented empirical evidence that it does not hold in general. Instead, they argued, comprehenders will choose the most typical interpretation. This is stated in their *Maximal Typicality Hypothesis* (yet another related hypothesis is due to Sabato & Winter, 2012).

3.2 COMPUTATIONAL COMPLEXITY OF QUANTIFIED RECIPROCALS

Szymanik (2009, 2010, 2016) studied the computational complexity of the decision problem (see sections 2.3 and 2.4) whether the truth conditions of quantified reciprocals are satisfied in finite models. Each of the possible readings (e.g. complete graph, path and pair readings) corresponds to a different decision problem. The input to the computational problem is some relational structure and the output is the decision whether it satisfies the truth conditions or not. Some of Szymanik's results, which are relevant below, are briefly discussed here. One is that path and pair readings of quantified reciprocals with tractable quantifiers as antecedents are tractable. For these cases, Szymanik sketches simple polynomial time algorithms.

Proposition 3:11. *Let Q be an UE quantifier of type $(1,1)$. If the decision problem associated with Q is in* **P**, *then $Ram_W(Q)$ and $Ram_I(Q)$ are in* **P**.

With regard to complete graph readings, the quantificational antecedent determines computational complexity. The complete graph reading of quantified reciprocals like 3:5 with *all* as antecedent can be decided in polynomial time.

Proposition 3:12. *$Ram_S(\textsc{All})$ is in* **P**.

The same holds for the quantificational antecedents *more than one*, *more than two*, *more than three*, etc. The complete graph reading of these reciprocals can be expressed in first order logic, as is also the case for reciprocals with *all* as antecedent (cf. 3:6-a). This implies polynomial time computability (e.g. Immerman, 1999). However, as Szymanik noticed, positing a different quantifier for each numeral

may not be cognitively realistic. It seems more plausible to assume one common semantic mechanism that computes the meanings and truth values for every *more than k* instead of distinct mechanisms for distinct *k*. To illustrate, it would be highly implausible that comprehenders would still have to learn the meaning of *more than 100* after they already had learned the meaning of *more than 99* and of the numeral *100*. To model this we introduce the counting quantifier $C^{\geq N}$:

Definition 3:13. $\mathrm{C}^{\geq N} := \{(M, R_1, R_2) : |R_1 \cap R_2| \geq |N|\}$.

As it turns out, verification of the complete graph reading is **NP**-complete if we consider a single semantic process that computes the decision problem for all possible *k*. This is a consequence of the fact that $Ram_S(C^{\geq N})$ is essentially a reformulation of the CLIQUE problem (see Definition 2:24 or Garey & Johnson, 1979 problem GT19).

Proposition 3:14. *$Ram_S(C^{\geq N})$ is* **NP***-complete.*

Furthermore, for reciprocals with proportional quantifiers like *most* (cf. Example 2:2, section 2.2), verification of the strongest meaning is also **NP**-complete. To proof this proposition, Szymanik describes a polynomial-time reduction of CLIQUE to $Ram_S(\textsc{Most})$ and other proportional quantifiers. Therefore, an effective solution would also imply an effective solution for the CLIQUE problem.

Proposition 3:15. *$Ram_S(\textsc{Most})$ is* **NP***-complete.*

Szymanik noted that the SMH and the PCT stand in conflict. While the SMH predicts a general preference for complete graph readings in case the quantificational antecedent is UE, the PCT predicts that complete graph readings should not be possible for reciprocals with proportional or counting quantifiers as their antecedents. Thus, combining the SMH and the PCT yields specific predictions about the interpretation of quantified reciprocals. Similar predictions could also be derived by combining the *Maximal Typicality Hypothesis* with the PCT. We can think of computational complexity as a filter acting on the possible meanings of reciprocal sentences: As long as all readings are computationally tractable, the logically strongest (or most typical) reading should be preferred. However, if the logically strongest (or most typical) reading is intractable, we should observe a shift towards tractable interpretations.

3.3 RELATED WORK

The PCT (Hypothesis 3:2) is an interesting claim because it severely restricts theories in cognitive science. Although it seems plausible that intractable problems cannot be mastered by cognitive agents with limited processing resources, matters may be more complicated. For example, a less strict variant of the Tractable Cognition Thesis was put forward by van Rooij (2008). This hypothesis is based on *fixed parameter tractability*. Fixed parameter tractable FPT problems are tractable as long as certain aspects, called parameters, of the problem instances are fixed. An intractable problem may well be FPT. Intuitively speaking, this is the case when the complexity of the problem only stems from specific aspects of the input. The FPT-Cognition hypothesis states that even functions that are intractable may be realistic cognitive functions as long as (i) certain parameters of the input can be identified that are responsible for the intractability and (ii) these usually stay within certain bounds.

3.3.1 *Theoretical work on computational complexity in semantics*

As stated in the introduction to the present chapter, computational complexity has not received much attention in semantics thus far. However, it has started to attract some attention lately. A recent review of this line of research was given by Isaac, Szymanik, and Verbrugge (2014, Section 4). We will shortly discuss some of the relevant work here. Besides quantified reciprocals, Szymanik (2009, 2010, 2016) also investigated the computational complexity of other kinds of quantified sentences. In a nutshell, the results of Szymanik's theoretical work in this area can be summarized as follows (see also Mostowski, 1998; Sevenster, 2006; Mostowski & Szymanik, 2007; Kontinen & Szymanik, 2008; Szymanik & Zajenkowski, 2010). Usually, the assumed meanings of natural language sentences with quantifiers are computationally tractable, but there are also a few intractable exceptions. Another example of intractable meanings, in addition to reciprocals, are so-called *Hintikka sentences*.

Besides the verification of quantifiers, two other areas of semantic competence have also been analyzed in terms of computational complexity. In both cases, the analysis has identified intractable problems that arise directly from standard assumptions in semantic theory. The first is anaphora resolution. Ristad (1993) has shown that, under stan-

dard linguistic assumptions, the computational problem of determining antecedents for anaphora, what he calls the 'anaphora problem', is **NP**-complete. Examples he discusses include sentences like *Mary said John saw himself* and *John hates his neighbors and so does Max*. Sentences of the latter kind turn out to introduce considerable complexity because of the contained ellipses. The second area is the satisfiablility problem for different fragments of English (e.g. Pratt-Hartmann, 2004; Pratt-Hartmann & Third, 2006; Pratt-Hartmann, 2008). This problem is also closely related to the computation of entailments. Pratt-Hartmann (2004) has shown that, assuming simplified versions of Montague grammar, determining satisfiability of a set of sentences easily becomes intractable. For example, we run into intractable problems if we enrich the traditional syllogistic fragment with relative clauses. For the latter fragment – which allows sentences like *every philosopher who is not a stoic is a cynic* – the decision problem is **NP**-complete. The approach of Pratt-Hartmann (2004) has been extended to more fragments of English by Pratt-Hartmann and Third (2006) and Pratt-Hartmann (2008). An overview can be found in Pratt-Hartmann (2010, Section 4).

The limited amount of existing theoretical work on computational complexity in semantics suffices to show that under common semantic assumptions several seemingly innocuous aspects of semantic competence involve computationally intractable problems. Under the PCT, this would, however, be unexpected. If intractable problems do indeed occur in semantic processing, it would pose a theoretical and empirical challenge to understand how people achieve solving these problems within reasonable time. One way of approaching this challenge would be in terms of the above-mentioned FPT-Cognition Hypothesis. In particular, we would have to look for aspects of the computational problems that have to be size-bounded in order for the processor to succeed within reasonable time. Another interesting perspective on this issue was proposed by Mostowski and Szymanik (2012). They argue on the basis of theoretical considerations that "everyday language" is restricted to meanings that can be expressed in the existential fragment of second-order logic, i.e. Σ^1_1. Since verification of Σ^1_1-sentences coincides with the problems in **NP** (Immerman, 1999), their thesis provides us – just like the PCT and the FPT-Cognition Hypothesis – with an upper bound on the computational complexity of meanings in everyday language. Similar to the FPT-Cognition Hypothesis, the Σ^1_1-thesis constitutes a relax-

ation of the PCT and would, in principle, allow for 'intractable' problems. To explain how people deal with these problems, Mostowski and Szymanik (2012) suggest that the most difficult verification problems within everyday language are verified indirectly by means of strategies that may involve inference from known facts or guessing strategies. To conclude our discussion of theoretical work on computational complexity in semantics, we want to mention one interesting application of the Σ^1_1-thesis. Kontinen and Szymanik provide a theorem which renders it implausible that a type-lifted, collective reading of *most* is definable in Σ^1_1 because, otherwise, the counting hierarchy would collapse at its second level. As (Szymanik, 2009, Section 5.5) pointed out, the Σ^1_1-thesis would, therefore, predict that this reading is impossible.

3.3.2 *Empirical work on computational complexity in semantics*

Empirical work on computational complexity in semantics is still scarce. Building upon the analysis of Pratt-Hartmann and Third (2006), Thorne (2012) has found initial but, as he notes himself, still inconclusive evidence that computational complexity is negatively correlated with how often sentences of certain types occur in a text corpus. He used automatic *deep semantic annotation* in order to map natural language sentences to different language fragments which differed in the computational complexity of the corresponding satisfiability problem. Moreover, evidence that points in the same direction was reported by Szymanik and Thorne (2015), who correlated the frequency of occurrence of quantified sentences with the computational complexity of the corresponding verification (or model-checking) problem.

A similar case as quantified reciprocals has been investigated by Gierasimczuk and Szymanik (2009). In particular, they studied the semantics of Hintikka sentences like the one in 3:16. Just like quantified reciprocals, these sentences have multiple potential readings of different logical strength. Furthermore, and also similar to the case of quantified reciprocals, it has been proposed in the semantic literature that the strongest reading of these sentences is the preferred one. Hintikka (1973) even claimed that this is their only possible reading. Again, the strongest reading corresponds to an **NP**-complete verification problem.

(3:16) More than 3 villagers and more than 5 townsmen hate each other

In line with the PCT, Gierasimczuk and Szymanik predicted that the strongest reading of Hintikka sentences is avoided due to its computational intractability. They used a sentence-picture verification and an inference task experiment to test whether Hintikka sentences allow for weaker readings than the strongest **NP**-complete one. Their findings indicate that this is in fact the case. They are thus in conflict with Hintikka's original hypothesis, but in accordance with the PCT.

What serves as the reference study for the following experiments is a first empirical study on the computational complexity of quantified reciprocals. In that study, Bott, Schlotterbeck, and Szymanik (2011) attempted to test the predictions from Szymanik (2009) experimentally. Their first experiment employed a picture completion task where participants had to construct situations that satisfied the preferred reading of German quantified reciprocals like DET *Punkte sind miteinander verbunden* ('DET dots are connected with each other'). They tested the quantifiers *alle* ('all'), *vier* ('four') and *die meisten* ('most'). They expected that, due to computational complexity, the choice of the quantificational antecedent should affect the preference for complete graph vs. path and pair readings. Reciprocals with *all* should receive a complete graph reading, but reciprocals with *most* and *four* should receive a path or pair reading. In contrast to what would be expected according to the SMH, participants overwhelmingly drew pictures not satisfying a complete graph reading. In the *all* condition, complete graph and path/pair interpretations were balanced. For the quantifiers *most* and *four*, path readings were strongly preferred, but complete graphs were hardly produced. Thus, the PCT received initial support. The second experiment employed a sentence-picture verification task in order to find out whether complete graph readings of intractable cardinal and proportional reciprocals were judged as impossible. Contrary to the expectations of Bott, Schlotterbeck, and Szymanik, complete graph readings were overwhelmingly rejected for all three types of quantifiers and there were no differences between them. At first glance, this lack of effect seems to contradict the PCT.

However, there are two potential problems with the study. Firstly, we cannot exclude the possibility of a floor effect covering potential differences because the complete graph readings were hardly avail-

able. Secondly, it is possible that the results were skewed towards general acceptance of path pictures and rejection of graph pictures because of the logical properties of the reciprocal relation *to be connected to*: it could have been interpreted as a transitive relation. As pointed out by Dalrymple et al. (1998), in reciprocals with a transitive relation the complete graph reading cannot be distinguished from the path reading. If this was the case in the study of Bott, Schlotterbeck, and Szymanik, nothing can be concluded from their picture verification results. To avoid these problems, a different – clearly intransitive – reciprocal relation was used in the present study. Next, we come to the empirical part of the present chapter which starts with a pretest that addressed the question whether *to be connected* really is interpreted transitively and compared it to another relation that seems clearly intransitive, intuitively.

3.4 PRETEST: THE RECIPROCAL RELATION

Example 3:10 (repeated below as 3:17) shows that the reciprocal relation (e.g. *to follow into the room* or *to know*) imposes important constraints on the available readings of reciprocal sentences. The predicate *to follow into the room* can only denote asymmetric relations; i.e., if *a* follows *b* into the room, *b* cannot follow *a*. The predicate *to know*, on the other hand, can also denote symmetric relations. Lexical restrictions like these can be encoded as *meaning postulates* (Carnap, 1952; Montague, 1973). In contrast to *know*, the lexical meaning of *follow into the room* is incompatible with the complete graph reading of quantified reciprocals.

(3:17) a. All students followed each other into the room.
b. All students know each other.

Another important property is transitivity. The predicate *to follow into the room* arguably denotes transitive relations; i.e., if *a* follows *b* into the room and *b* follows *c* into the room, then *a* does also follow *c* into the room. In case the reciprocal relation is transitive, models containing a path of the relevant size are always compatible with both the path and the complete graph reading of quantified reciprocals. Take, for example, the picture in Figure Xb. If we assume that the (partly) depicted relation is transitive, this picture is compatible with the path as well as the complete graph reading of quantified reciprocals. For

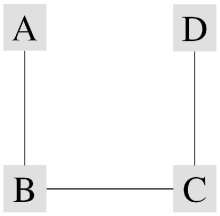


Figure XI. Diagram used to test for (in)transitivity of the relations denoted by *to be connected to* and *to be directly connected to*.

our test case, it is thus crucial to choose a relation that is not biased towards a transitive (or asymmetric) relation.

3.4.1 *Methods*

In a pretest, two candidate relations were compared. These were *to be connected* and *to be directly connected*. The goal was to find a relation that is interpreted intransitively. The pretest was a sentence-picture verification experiment with non-quantificational reciprocal sentences. The picture in Figure XI together with one of the sentences in 3:18 was presented to 80 native German speakers recruited at the University of Tübingen. Each of them provided only one truth value judgment. This way, it was ensured that the data were unbiased because there was no earlier exposure to another condition. Each sentence was tested in a subgroup of 20 participants. The test was carried out using paper and pencil.

(3:18) a. A und D sind miteinander verbunden.
A and D are with-one-other connected.
'A and D are connected to each other.'

b. A und D sind nicht miteinander verbunden.
A and D are not with-one-other connected.
'A and D are not connected to each other.'

c. A und D sind direkt miteinander verbunden.
A and D are directly with-one-other connected.
'A and D are directly connected with each other.'

d. A und D sind nicht direkt miteinander verbunden.
A and D are not directly with-one-other connected.
'A and D are not directly connected with each other.'

The proportion of *yes, true* judgments in reaction to sentence 3:18-a were taken as an estimate of how easily *to be connected* can be interpreted transitively. By contrast, judgments of the negated sentence in 3:18-b were informative about how easily the relation *to be connected* can be understood intransitively. The same holds with regard to *to be directly connected*: 3:18-c probed for the transitive interpretation and 3:18-d for the intransitive interpretation. In order to decide whether the transitive or intransitive interpretations are possible, it was statistically analyzed whether an observed level of acceptance significantly differed from complete rejection, i.e. null acceptance. One trial in both of the conditions in 3:18-b and 3:18-c was excluded from the statistical analysis because no judgment had been given.

3.4.2 *Results and discussion*

Judgments were distributed as follows: 3:18-a was accepted in 80% of the cases, 3:18-b was accepted in 42%, 3:18-c in 11% and 3:18-d in 80% of the cases. Fisher's exact test revealed that the proportions of *yes, true* judgments in the conditions 3:18-a, 3:18-b and 3:18-d were significantly different from 0% ($p < .01$, in all cases), whereas condition 3:18-c didn't differ significantly from 0% ($p = .487$). With respect to *to be connected to* the results show that the relation is in fact ambiguous. There was a preference for the transitive reading as indicated by the higher proportion of *yes, true* judgments of sentence 3:18-a than 3:18-b. This shows that the objections raised in connection with the study by Bott, Schlotterbeck, and Szymanik (2011) were well taken and *to be connected to* can indeed be interpreted transitively. Regarding *to be directly connected to*, there is a strong preference to interpret this relation intransitively as indicated by general rejection of 3:18-c. Furthermore, it seems to be clearly symmetric on intuitive grounds: if *a* is directly connected to *b*, *b* certainly is to *a*. Therefore, this relation fulfills all semantic requirements to be used in the two main experiments.

3.5 EXPERIMENT 1: PICTURE COMPLETION

In order to investigate whether the computational complexity of sentence verification affects comprehension processes, interpretation preferences were surveyed in a picture completion experiment. Partici-

pants were asked to draw connections between dots in a picture in such a way that the picture matched their preferred interpretation. As outlined above, the complete graph readings of the following sentences differ in computational complexity.

(3:19) a. All of the dots are directly connected to each other.
b. Four of the dots are directly connected to each other.
c. Most of the dots are directly connected to each other.

The complete graph reading of the reciprocal with *all* is computationally tractable whereas it is intractable for the other two quantificational antecedents. Under the above-mentioned assumption that intractable meanings are generally avoided by the comprehension system (Avoidance Strategies A and B, Hypotheses 3:3 and 3:4), the PCT and the SMH make conflicting predictions. The SMH predicts that complete graph readings are generally preferred whereas the PCT amended with Avoidance Strategy A or B only allows for complete graph readings if they are computationally tractable. How to apply the *maximal typicality hypothesis* of Kerem et al. (2011) to these cases is not obvious because it is unclear what the most typical scenario for *to be directly connected to* looks like. Probably, the different readings do not differ in typicality and should hence be equally likely. Again, computational complexity could act as a filter on the possible interpretations.

3.5.1 *Methods*

3.5.1.1 *Materials, Procedure and Participants*

Twenty-two native German speakers (mean age 22.2 years; 13 female) participated in the experiment for course credit in a third semester syntax class. Participants were naïve to the purpose of the study. The test was conducted using paper and pencil. Participants received a series of sentences, each paired with a picture of yet unconnected dots. Their task was to connect the dots in a way that the resulting picture matched their interpretation of the sentence. The following German sentences were used.

(3:20) a. Alle Punkte sind direkt miteinander verbunden.
All dots are directly with-one-other connected.
'All dots are directly connected with each other.

b. Vier Punkte sind direkt miteinander verbunden.
Four dots are directly with-one-other connected.
'Four dots are directly connected with each other.'

c. Die meisten Punkte sind direkt miteinander
The most dots are directly with-one-other
verbunden.
connected.
'Most dots are directly connected with each other.'

Sentences with *alle* ('all') were always paired with a picture containing four dots, whereas *vier* ('four') and *die meisten* ('most') had pictures with seven dots. There were fifteen experimental trials, five per condition, which differed with respect to the arrangement of the dots. In addition, 48 filler sentences were included. These were of two types. Half of them clearly required a complete graph (e.g. *only one dot is not directly connected to all of the other dots*). The other half was only consistent with a path (e.g. *six dots are connected to each other and form the letter S*). Four pseudo-randomized lists were constructed, in which two adjacent items were always separated by at least two fillers and each condition was as often preceded by a complete graph filler as it was by a path filler. This was done to prevent biases towards either complete graph or path interpretations. Two of the four lists were inverted versions of the other two lists. This way, effects of presentation order were controlled for.

The completed pictures were annotated with respect to the chosen interpretation. We classified a picture to represent a complete graph reading if it satisfied the truth conditions in 3:7-a. Pictures were taken to indicate a path reading if a sufficiently large subgraph was connected by a continuous path (3:7-b), but there was no complete graph connecting the nodes. Finally, it was taken to indicate a pair reading if the required number of nodes were interconnected (3:7-c), but there was no path connecting them all. One participant did only return 14 out of the 15 experimental trials because one page of the questionnaire was missing. Thus, 329 productions were analyzed in total.

3.5.1.2 *Predictions*

According to the SMH complete graph readings are generally preferred for reciprocals with UE quantificational antecedents. Thus, it predicts high proportions of complete graph readings across the three tested quantifiers (granted that *four* has an UE interpretation, namely

as *four or more*). The PCT in combination with one of the above-mentioned avoidance strategies predicts that complete graph readings are dispreferred if they are computationally intractable. Thus, reciprocals with *four* or *most* as antecedent should receive a low proportions of complete graph interpretations. As far as these hypotheses are compatible, there combined prediction would be that the *all* condition overwhelmingly receives complete graph interpretations whereas the conditions with *four* or *most* receive a path or a pair reading in the majority of the cases.

3.5.1.3 *Statistical Analysis*

In order to test the predictions logit mixed effects models were computed (see Jäger, 2008 for a plea to use this kind of model to analyze categorical data and Gelman & Hill, 2007 for an introduction). The model included the factor *quantificational antecedent* (levels: *all*, *four* and *most*) as fixed effect and by-participant random intercepts and random slopes (see Barr, Levy, Scheepers, & Tily, 2013 Bates, Kliegl, Vasishth, & Baayen, 2015 for discussion of random effect structures). Random effects of items were not included because participants always received the same sentence within each condition, the only thing that varied were the dot pictures. The models were computed using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) of the R software (R Core Team, 2016). As test statistic, Wald's *Z*, provided by lme4 package, was used. Dummy coding was used in order to compare factor levels *all* and *most* to *four*, which served as reference level.

3.5.2 *Results*

Overall, complete graphs (57.3%) and paths (41.8%) were produced rather frequently whereas pairs were only produced two times, both in the *most*-condition (0.6%). The proportions of complete graphs in each of the three conditions are provided in Table XII. In the *all*-condition, participants chose complete graph readings 74.6% of the time. By contrast, in the *four*-condition, there were only 53.6% complete graphs. The number of complete graphs was even lower in the *most*-condition with only 44.0%.

The estimated parameters for the fixed effects are presented in Table XII. The logit mixed effects model analysis revealed that *all*

Table XII

Descriptive and inferential statistics of Exp. 1.

(a) *proportions of complete graphs*

	proportion
all	.746
four	.536
most	.440

(b) *parameters of the logit mixed effects model*

	estimate	Z	p
(intercept)	0.08	0.09	.928
all vs. *four*	2.53	4.03	$< .001$
most vs. *four*	-0.72	-1.20	.230

led to a significantly higher proportion of complete graphs than *four* as reflected by the significant fixed-effect of *all*. However, the 9.4% difference between *most* and *four* was not significant.

3.5.3 *Discussion*

The preference for complete graph readings in the *all*-condition is in line with the SMH. Furthermore, the lower proportions of complete graph readings for *most* and *four* than for *all* reciprocals matched the predictions of the PCT. The complete graph interpretations of both *most* and *four* reciprocals constitute intractable meanings and – in line with our considerations in the introduction to the present chapter – should hence be avoided. This result is particularly interesting because the experiment did not involve solving an intractable verification problem, but instead focused on comprehension processes. Constructing a complete graph is not intractable even for *most* and *four* reciprocals since any large enough subset of vertices could be randomly chosen.[4] The observed effect can thus be taken as an indication that the comprehension system in fact tries to avoid intractable meanings.

In the introduction of this chapter, we sketched two alternative scenarios in order to explain how avoidance of intractable meanings can be achieved. According to Avoidance Strategy B (Hypothesis 3:3), intractable complete graph readings should never occur. This is because the comprehension system is assumed to be restricted in a way that it cannot represent intractable natural language meanings. The data do not fit this theoretical option. For intractable antecedents, we observed complete graphs in approximately half of the cases. The data are, however, fully consistent with Avoidance Strategy A (Hy-

[4] We would like to thank Iris van Rooij for pointing this out.

pothesis 3:4). There, we assumed that the comprehension system would avoid intractable meanings on the basis of prior experience. It is, therefore, fully expected that intractable meanings occur in some of the trials, in case comprehenders fail to identify the computational intractability of a meaning.

3.6 EXPERIMENT 2: SENTENCE-PICTURE VERIFICATION

The second experiment tested verification of quantified reciprocals which were presented together with pictures that disambiguated complete graph from path readings. To achieve clear disambiguation, different quantificational antecedents had to be used than in the previous experiment. This is because the quantifiers tested in the previous experiment were all UE and, therefore, complete graphs are also compatible with a path reading. In the present experiment, we used reciprocals with *all but one* and *exactly k*, as in 3:21. *All but one* and *exactly k* are clearly non-monotone (see Definition 2:8), and hence none of the readings entails the others (see Dalrymple et al., 1998, p. 206, proposition 4).

(3:21) a. Alle Punkte bis auf einen sind direkt miteinander
All dots up to one are directly with-one-other
verbunden.
connected.
'All but one dots are directly connected to each other.'

b. Genau drei/fünf Punkte sind direkt miteinander
Exactly three/five dots are directly with-one-other
verbunden.
connected.
'Exactly three/five dots are directly connected to each other.'

In order to capture the possible meanings of quantified reciprocals with non-monotone quantificational antecedents, we have to use different formulas than in 3:7 above (section 3.1). Otherwise, complete graph readings would still entail path readings, which is what we want to avoid because it does not correspond to the intuitive interpretations of these sentences. Szymanik (2010) restricted the operators Ram_S, Ram_I and Ram_W to upward entailing quantifiers. Dalrymple et al. (1998) suggested a somewhat involved uniform treatment for all type $(1,1)$ quantifiers that can be used regardless of their monotonicity. For the present purpose, it is, however, sufficient to formalize the relevant readings as follows. The path readings of *all-but-one* and

exactly-k reciprocals are formalized as in 3:22-a and 3:22-b, respectively. The complete graph readings are formalized analogously, as in 3:22-c and 3:22-d. The result is truth-conditionally equivalent to the suggestion of Dalrymple et al.

(3:22) a. $Ram_I(\textsc{AtLeastAllButOne}) \wedge \neg Ram_I(\textsc{All})$
b. $Ram_I(C^{\geq N}) \wedge \neg Ram_I(C^{\geq M})$
c. $Ram_S(\textsc{AtLeastAllButOne}) \wedge \neg Ram_S(\textsc{All})$
d. $Ram_S(C^{\geq N}) \wedge \neg Ram_S(C^{\geq M})$

The quantifier AtLeastAllButOne is defined as $\{\mathcal{M} : 1 \geq |A \setminus B|\}$. The sets N and M in 3:22-b and 3:22-c are meant to be such that $|M| = |N| + 1$. Conjunction and negation of GQs are defined as usual (e.g. Szymanik, 2016, p. 30).

The complexity results of Szymanik (2010) do not apply directly to the conjunctions in 3:22. Note, however, that these are tractable if their conjuncts or the negations thereof are (cf. Immerman, 1999). In combination with Szymanik's results, this implies that 3:22-a–c are tractable.

Proposition 3:23. *The decision problem associated with the GQs in 3:22-a–c are in* **P**

Crucially, although intuitively more complex than simple *all*, the complete graph reading of *all-but-one* reciprocals is also in **P**. A brute force algorithm requires approximately n-times as many steps as an algorithm to verify *all* reciprocals. In order to verify a model of cardinality n, at most the n subsets of cardinality $n - 1$ have to be considered.

By contrast, the complete graph reading of *exactly-k* reciprocals is intractable since it is equivalent to the MAXCLIQUE problem (see e.g. Garey & Johnson, 1979, p. 164 or Definition 2:25, in section 2.4) which is **NP**-hard (see proposition 2:26). This problem consists in deciding whether the maximum complete subgraph in a graph is of size k. In particular, *exactly-k* reciprocals are intractable because k is not a constant, but a variable (see Section 3.3 and Szymanik, 2010, for discussion). In the present experiment *all-but-one* was kept constant and *exactly k* was presented as *exactly three* and *exactly five*.

Proposition 3:24. *Let N and M be two finite sets such that* $|M| = |N| + 1$. $Ram_S(C^{\geq N}) \wedge \neg Ram_S(C^{\geq M})$ *is* **NP** *-hard.*

We assume that a quantifier like *all but one* is represented as an atomic lexical item, i.e. as an idiom, and is thus different from *exactly*

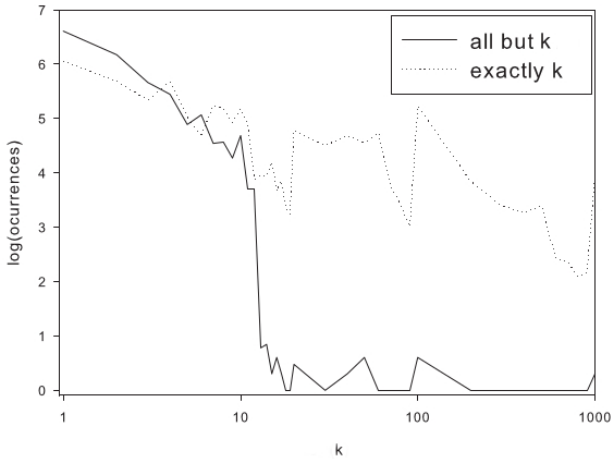


Figure XIII. Absolute frequencies of German *all but k* and *exactly k* in a Google search. Note: values were log-transformed with a base of 10.

three and *exactly five* which are instances of *exactly k* with a compositionally derived meaning. Since this assumption is crucial for the present experiment, it was qualified by a small corpus study in which the absolute frequencies of strings of the form *alle bis auf k* (‘*all but k*’) and *genau k* (‘*exactly k*’) were googled. The search covered the following values of *k*: 1 to 20 in steps of 1, 20 to 100 in steps of 10 and 100 to 1000 in steps of 100. The results are shown in Figure XIII. In line with our assumptions, there were hardly any instances of *all but k* for $k \geq 13$ (51 hits in total), whereas *exactly k* had more than 480,000 instances for numbers between 13 and 1000. The two types of quantifiers in 3:21-a and 3:21-b thus are, in fact, fairly different from each other. *Exactly k* is compatible with variable *k*, but *all but k* seems to be limited to a small set of constants. To assume that *all but one* is an idiomatic expression may be an overstatement. However, it seems reasonable to assume that, in contrast to *exactly k*, no single mechanism that computes the meanings and truth-values of all the different *all but k* is readily available.

The sentences in 3:21 were paired with diagrams disambiguating towards the complete graph or the path reading. Sample diagrams are shown in Figure XIVa/e and XIVb/f, respectively. As for complete graph pictures, the PCT let us expect lower acceptance of 3:21-b than of 3:21-a. On the strictest interpretation of the PCT, we may even expect null acceptance of complete graphs given 3:21-b. In order to be able to find out whether the complete graph readings of

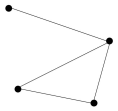
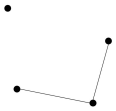
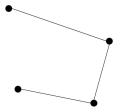
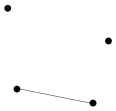
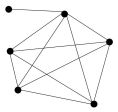
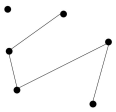
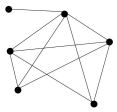
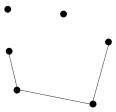
$n = 4$			
			
(a) complete subgraph	*(b)* continuous path	*(c)* false baseline for complete subgraphs	*(d)* false baseline for continuous paths
$n = 6$			
			
(e) complete subgraph	*(f)* continuous path	*(g)* false baseline for complete subgraphs	*(h)* false baseline for continuous paths

Figure XIV. Sample diagrams presented in the picture verification experiment. The upper row presents graphs and paths with four dots. Graphs and paths with six dots are shown in the bottom row.

3:21-b are possible at all, the sentences in 3:21 were also paired with false diagrams which served as baseline controls (see Figure XIVc/g). The controls differed minimally from the complete graph pictures in that a single line was removed from the completely connected subset. If the complete graph reading is possible, we should observe more *yes, true* judgments in the complete graph conditions than in the false controls. The true and false path diagrams (e.g. XIVb/f and XIVd/h, respectively) were included to control whether participants did understand the quantificational antecedents.

Additionally, the size of the models was manipulated. It is conceivable that people can verify intractable complete graph readings given small graphs, but fail to do so for larger ones due to computational complexity. Therefore, besides having the two quantifiers *all but one* and *exactly k*, another manipulation consisted in the size of the graphs, i.e. the number of vertices. Small graphs always contained four dots (see Figure XIVa-d) whereas large graphs contained six dots (see Figure XIVe-h). This way, it was possible to keep the quantifier *all but one* constant and compare it to *exactly k* with variable *k*. *Exactly k* was instantiated as *exactly three* and *exactly five*, respectively. In to-

tal, this yielded 16 conditions according to a 2 (*quantifier*) × 4 (*picture type*) × 2 (*graph size*) factorial design.

3.6.1 *Methods*

3.6.1.1 *Materials, participants and procedure*

Nine experimental items in 16 conditions like the sample item in 3:21 and Figure XIV were constructed. A Latin square was used to make sure that each picture only appeared once for each participant, but that the same picture appeared with both quantifier types. [5] Each participant provided three judgments per condition resulting in a total of 48 experimental trials. Sixty-Six filler trials were added to the experiment and, for each participant, trials were individually randomized.

Thirty-four native German speakers (mean age: 27.5y, 20 female) read quantified reciprocal sentences on a computer screen. After reading the sentence, they had to press a button which made the sentence disappear and a dot picture appear for which they had to provide a truth-value judgment.

3.6.1.2 *Predictions*

Under the strictest reading of the PCT and likewise under Avoidance Strategy B (Hypothesis 3:4), complete graph interpretations of reciprocals with *exactly k* should not be possible. In contrast, *all-but-one* reciprocals should allow for complete graph interpretations if our assumptions concerning their meaning are correct. In combination, this predicts that acceptance of complete graph diagrams (e.g. panels a and e of Figure XIV) and their respective false controls (e.g. panels c and g, respectively) do not differ in conditions with *exactly-k* reciprocals, but that there is a clear difference within conditions with *all-but-one* reciprocals. Path interpretations are always tractable and therefore participants are expected to be able to distinguish correct path diagrams from their corresponding false controls (e.g. panels

[5] The experiment included eight more conditions. These were reciprocal sentences with the antecedent *most*. Originally, it was expected that *most* would trigger the scalar implicature *not all*, which would have allowed us to compare it to *all but one* and *exactly k*. The results indicated that this was not the case, or rather, that the implicature could easily be canceled. In the *false graph* condition with four dots, for instance, more than 60% acceptance was observed, making it impossible to properly analyze *most*-reciprocals. Therefore, these conditions are not reported here. For future research, it would be interesting to test non-monotone quantifiers of the type *most but not all* or *exactly half* which would obviously constitute excellent test cases for the PCT.

panels b and f vs. panels d and h, respectively), regardless of the quantificational antecedent. This is expected to lead to higher acceptance of the former as compared to the latter type of pictures.

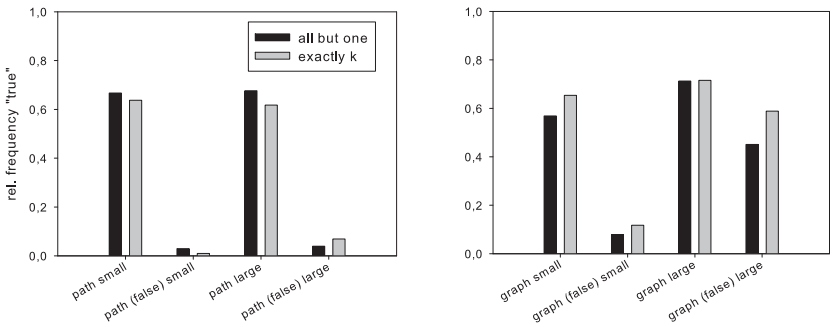
If we assume, in line with Avoidance Strategy A (Hypothesis 3:3), that intractable interpretations are not ruled out in general, but lead to substantial difficulty during verification, a more graded pattern of results is expected. In particular, the acceptance of complete graph pictures in conditions with *exactly-k* reciprocals is expected to be higher than the acceptance of the false baseline controls. In addition, participants are, however, also expected to give a substantial amount of erroneous responses when evaluating *exactly-k* reciprocals against complete graph diagrams. In conditions with *all-but-one* reciprocals, the proportions of erroneous responses is expected to be comparatively low. This predicts a two-way interaction of the factors *picture type* (true vs. false) and *quantifier* if we restrict ourselves to the complete graph conditions, i.e. conditions with diagrams as shown in panels a/c and e/g of Figure XIV. Within the continuous path conditions (e.g. panels b/d and f/h), no such interaction is expected.

Under Avoidance Strategy A, where intractable interpretations are not ruled out in general, the number of dots in the pictures may also affect experimental results. Specifically, we may observe little or no difference between the two quantificational antecedents in conditions with only four dots but find results that clearly reflect the just described interaction in conditions with six dots. Together, this would lead to a significant three-way interaction between the factors *quantifier*, *picture type* and *graph size* within the complete graph conditions whereas the *quantifier* is not expected to have any effect within the continuous path conditions.

3.6.1.3 *Statistical analysis*

The path and complete graph conditions were analyzed separately, performing logit mixed effects model analyses on proportions of acceptance. The fixed effects of *quantifier* (levels: *all but one* and *exactly k*), *graph size* (levels: *four* and *six*) and *truth* (levels: *true* and *false*) and their interactions were included in the models as well as by-participant random intercepts and random slopes. Random effects of items were not included because, as in Experiment 1, the experimental sentences were the same across conditions. [6] The purpose of these

[6] No interactions were included into the random slopes of subjects because these models failed to converge.



(a) continuous paths *(b)* complete graphs

Figure XV. Proportions of acceptance in the picture verification experiment (*small*: pictures with 4 dots; *large*: pictures with 6 dots). Panel a shows judgments for the path conditions; panel b for the complete graph conditions.

analyses was to test whether *quantifier* – possibly in interaction with *graph size* – affected the proportions of errors. Proportions of errors are reflected in the frequencies of how often the true conditions are accepted relative to the false conditions. Thus, the statistical analysis tested for effects involving *quantifier*×*truth*-interactions. More specifically, it was tested whether the effect of *truth* is larger within *exactly-k* than within *all-but-one* conditions, as predicted.

In order to break down significant interactions, it was planned to compute logit mixed effects models on subsets of the data. This is relevant in the discussion of the complete graph conditions below, where we report separate models for the true and the false complete graph conditions.

3.6.2 *Results*

Proportions of acceptance are depicted in Figure XV. In the following, the results of the path conditions shown in Panel XVa are described first. The complete graph conditions in Panel XVb are described afterwards. The estimated parameters for the fixed effects are presented in Table XVI.

PATH DIAGRAMS: The path reading was generally accepted (true path conditions: mean acceptance of 67.7%) and led to almost no errors (false path conditions: mean acceptance of 4.2%) with both

Table XVI

Estimates for fixed effects, Wald's Z statistics and p-values of the logit mixed effects analyses of complete graphs in Exp. 2.

		estimate	Z	p
Global analysis				
	(intercept)	-0.26	-0.70	0.48
	quantifier	0.81	2.00	$<.05$
	truth	2.07	4.31	$<.01$
	graph size	-2.34	-4.49	$<.01$
	quantifier×*truth*	-0.99	-1.85	0.06
	quantifier×*size*	-0.92	-1.42	0.16
	truth×*size*	0.91	1.52	0.13
	quantifier×*truth*×*size*	1.60	1.92	0.06
True conditions				
	(intercept)	1.78	3.26	$<.01$
	quantifier	-0.25	-0.48	0.63
	graph size	-1.34	-3.05	$<.01$
	quantifier×*size*	0.81	1.51	0.13
False conditions				
	(intercept)	-0.30	-0.85	0.39
	quantifier	0.78	2.29	$<.05$
	graph size	-2.51	-4.66	$<.01$
	quantifier×*size*	-0.72	-1.15	0.25

quantifiers and graph sizes. The statistical analysis revealed that only the main effect of *truth* was significant ($p < .01$). Thus, with respect to their tractable interpretation, both quantificational antecedents behaved similarly.

COMPLETE GRAPH DIAGRAMS: As expected, the true complete graph diagrams (Figure XIVa/e) were accepted across the board (66.3%). Furthermore, true complete graph diagrams were accepted significantly more often than false ones ($p < .01$). In the false complete graph conditions (Figure XIVc/g), there were, however, clear differences between diagrams with four and six dots. In the former conditions, participants made relatively few errors (9.9%) whereas proportions of errors were substantially higher in the latter conditions (52%).

This led to a reliable effect of *graph size* ($p < .01$). Moreover, *exactly-k* reciprocals were accepted more often than *all-but-one* reciprocals as reflected by a significant main effect of *quantifier* ($p < .05$). Finally, two interactions were marginally significant. Firstly, the two-way interaction between *truth* and *quantifier* was marginally signifi-

cant ($p = .06$), reflecting the fact that the difference in acceptance rates between the two quantificational antecedents was bigger in the false than in the true complete graph conditions. Secondly, the three-way interaction was marginally significant ($p = .06$).

In order to break down the three-way interaction, separate mixed effects models for the true and false complete graph conditions were computed. In the false complete graph conditions, the main effect of graph size was significant ($p < .01$) because participants made more errors in graphs with six dots than in graphs with four dots. Furthermore, *exactly k* led to more erroneous responses than *all but one*. This effect was consistent across graph sizes as indicated by a significant main effect of *quantifier* ($p < .05$) and a non-significant *graph size*×*quantifier* interaction ($p = .25$). Thus, *exactly k* lead to more errors than *all but one* consistently across both graph sizes. In the separate analysis of the true complete graph conditions, the only significant effect was the main effect of *graph size* ($p < .01$). This effect indicates that acceptance rates were higher in the larger graphs than in small graphs.

3.6.3 *Discussion*

The results show that quantified reciprocals allow for complete graph readings. The true complete graph conditions were overwhelmingly accepted and participants were able to distinguish them from the false controls. As opposed to the predictions made by the strictest reading of the PCT, this was the case for both *all-but-one* and *exactly-k* reciprocals. In the false conditions, we did, however, find indications of effects of semantic complexity. In these conditions, proportions of errors were higher for intractable *exactly-k*- than for tractable *all-but-one* reciprocals. This is what we may expect according to a less strict interpretation of the PCT. Based on this interpretation, it was expected that an increase in graph size would lead to a steep increase in error rates. In fact, in the false conditions, graph size did lead to an increase in the number of erroneous responses. However, we expected a larger effect for intractable *exactly-k* than for tractable *all-but-one* reciprocals. Despite this trend being present numerically, it was not reliable. It, thus, remains an open question, whether for larger graph sizes the effect of graph size on the verification of *exactly-k* reciprocals is stronger than it is on the verification of *all-but-one* reciprocals.

Why did computational complexity affect only the false conditions? This might be due to the specific verification procedures participants were able to use. In the true complete graph conditions, the complete subgraph was visually salient and could thus be immediately identified (Figure XIVa/e). This, in turn, may have reduced the computational resource demands of the decision problem participants had to solve. A reasonable procedure to verify the complete graph reading of the *exactly-k* reciprocals in 3:22-d consists of two subroutines. In a first step, participants could decide whether there is a complete subgraph of size *k*. In a second step, they would then have to ensure that there is no complete subgraph of size *k+1*. The first step would consist in solving the CLIQUE problem. Once participants have identified the relevant subgraph, the first step can be completed in polynomial time because the completely connected subgraph is a *certificate* for the CLIQUE problem (the term certificate was discussed in connection with the class **NP** in Section 2.4). In the picture material we used, the second step would be trivial because there was only one vertex in addition to the completely connected subgraph (see Figure XIVa/e). As a consequence, participants really had to solve an intractable problem only in the false conditions with one edge removed from the graph. Apparently, this was still possible when the relevant subgraph consisted of only three dots, but already started to exceed cognitive capacities when it consisted of five dots.

3.7 CONCLUSIONS

We started with hypotheses that made rather strong predictions. Linguistic work on reciprocal sentences by Dalrymple et al. (1998), who introduced the SMH, led us to expect that ambiguous reciprocals should receive their logically strongest interpretation. Therefore, a complete graph interpretation should be chosen for reciprocals with UE antecedents. Szymanik (2010), on the other hand, employed the PCT to predict shifts in interpretation in case the quantified reciprocal has an intractable complete graph reading.

For tractable reciprocals with *all* as antecedent, the predictions of the SMH were borne out. In the picture completion experiment, participants overwhelmingly drew complete graphs in reaction to reciprocal sentences with the antecedent *all*. Furthermore, results of the picture completion experiment were in line with the predictions of the PCT for comprehension. The quantificational antecedent influ-

enced interpretation preferences. For *most* and *four* reciprocals, participants produced fewer complete graphs than they did for *all* reciprocals. This indicates that comprehenders avoid intractable meanings.

In the introduction, we have outlined two scenarios in order to explain how the sentence processor can avoid interpretations involving intractable verification problems. In the first scenario (Avoidance Strategy A) we assumed that, on the basis of earlier exposure, comprehenders will stay away from interpretations that have in the past turned out to be too complex to verify. To do so, however, they have to be able to decide beforehand whether a meaning leads to an intractable verification problem or not. The decision may depend on simple heuristics which can sometimes fail and which may also differ between quantificational antecedents. One conceivable example of such a heuristic would be to memorize types of linguistic expressions that do in certain constructions lead to intractable verification problems. The second scenario outlined in the introduction (Avoidance Strategy B) rested on the assumption that the language faculty is set up in such a way that comprehenders never face an intractable verification problem, e.g., by constraining the formal properties of the language in which humans may internally represent natural language meanings.

These two scenarios gave rise to slightly different predictions. In the first scenario, it would be expected that – at least sometimes – comprehenders do not notice that they are running into an intractable problem. It would, therefore, be expected that intractable meanings do occur, although less frequently than their tractable counterparts. Furthermore, intractable reciprocals with different quantificational antecedents may differ in how easy it is to identify that verification of the complete graph reading poses an intractable problem. Therefore, the resulting shift in meaning may be stronger with one quantificational antecedent than with another one. Thus, the first scenario let us expect gradient shifts in preference. According to the second scenario, intractable interpretations should not occur and the shift should thus be absolute. The findings of the picture completion experiment are only compatible with the first scenario. The drop of graph readings was far from absolute. Instead, we still observed approximately 50% complete graph pictures for intractable antecedents. Even though we have to be careful not to over-interpret this result – due to the logical entailments between readings — it seems highly plausible that participants had complete graph interpretations in mind when drawing

a complete graph picture. Complete graphs required drawing more connections than the alternative readings and we find it implausible that participants would systematically draw more connections than necessary.

The results of the picture verification task experiment complement the picture. Intractable readings were clearly available to our participants. As far as our assumptions concerning *exactly-k* reciprocals are correct, this result provides evidence against the PCT under its strictest reading. Nevertheless, we observed effects of semantic complexity. Participants performed close to chance level when they had to reject pictures which did not satisfy the complete graph reading because of one missing connection. In these conditions, the number of errors was slightly higher for intractable *exactly-k* than for tractable *all-but-one* reciprocals.

A word of caution is in order here. Based on the limited amount of data presented here, we cannot entirely rule out the possibility that participants derived approximate interpretations which differ from the assumed truth conditions. Apart from this caveat, the findings from the picture verification experiment indicate that intractable interpretations are not ruled out in general, but computational complexity effects only emerge with certain problem instantiations. The results of the experiment indicate that comprehenders (at least sometimes) arrive at intractable interpretations during comprehension, but depend on strategies that reduce complexity in a verification setting. In particular, we speculated that the salience of the completely connected subgraph in a subset of our stimulus materials provided a cue to guess the relevant subgraph via visual pattern recognition. This would considerably simplify the verification task and could, on the other hand, explain why error rates increased up to chance level when the number of vertices was slightly increased and visual salience provided no cue. Since the presence of visual cues is not guaranteed, this line of reasoning suggests that the verification of intractable complete graph readings does, in fact, exceed cognitive capacities. It would be interesting to systematically manipulate the visual salience of the relevant subgraph in order to test whether this explanation is on the right track. As exemplified by the work of Pietroski et al. (2009), Lidz et al. (2011) and Tomaszewicz (2013), among others, studying the concrete verification procedures that are used for verification of a sentence's truth conditions is interesting in its own right.

To conclude, Avoidance Strategy B seems to provide the best explanation for our findings. That is, despite a general tendency to avoid them, the comprehension system sometimes derives intractable interpretations. These interpretations do, however, lead to severe difficulties during verification and can only be properly dealt with by relying on guessing strategies, or other indirect methods. Taken together, our findings suggest that restricting the semantic processor to verification problems in **P** is too strict. This interpretation is in line with considerations of Ristad (1993) and Mostowski and Szymanik (2012). The latter also argue that intractable verification problems are indirectly verified employing guessing strategies. [7] A relaxation of the PCT has also been suggested by van Rooij (2008) with regard to other cognitive capacities. A potentially interesting route to pursue would be to apply an analysis as outlined by van Rooij (2008) based on fixed parameter tractability. This, however, has to be left to future research.

[7] In connection to this consultation of Halpern (2003) might be instructive.

4

EXISTING PROCESSING MODELS OF QUANTIFIER VERIFICATION: THE AUTOMATA MODEL, INTERFACE TRANSPARENCY AND THE APPROXIMATE NUMBER SYSTEM

In the previous chapter, abstract considerations of computational complexity enabled us to predict some aspects of the comprehension and verification of quantified sentences that would not have been expected otherwise. But at the same time, it became evident that a dichotomous distinction between computationally tractable and intractable problems does not suffice to account for all aspects of the obtained experimental data. It seems that the specific processes at play have to be studied for this. Especially since candidates for intractable natural language meanings, in the above sense, are rare (e.g. Szymanik, 2016, parts 2 & 3), it seems important to understand which mechanisms are employed to tackle the tractable ones.

That good progress can be made in studying the processes underlying the truth evaluation of quantified sentences is reflected in an increasing number of publications during the last decade or so. The present chapter discusses two common approaches to this issue. The first is the *semantic automata model* (originally proposed by van Benthem, 1986): a computational model of quantifier verification and falsification that allows us to study the processing requirements of individual quantifiers or subclasses thereof. The second can be summarized under the idea of *interface transparency* (explicitly formulated by Lidz et al., 2011): the hypothesis that the compositional encoding of truth conditions to some degree determines the verification process associated with a quantified sentence.

4.1 THE AUTOMATA MODEL: THEORY AND EXPERIMENT

A link between the semantics of quantifiers and computation was proposed by van Benthem (1986) (referring to informal ideas of Suppes) and further developed, for example, by Mostowski (1998), Szymanik (2009), Steinert-Threlkeld and Icard (2013) and Steinert-Threlkeld (2016). In a nutshell, the basic idea is, again, to conceive of quantifiers as *languages* or, equivalently, *decision problems* (in the sense of section 2.3, Definition 2:14) and study these from a computational perspective. A crucial difference to above, is that all the quantifiers discussed below correspond to functions that are computable in polynomial time. But still, there are differences in computational recourses that are needed to recognize the languages associated with these quantifiers. Furthermore, we will see below that some of these distinctions also have psychological relevance, as was confirmed in psycholinguistic experiments.

4.1.1 *Theory*

The semantic automata model is best explained using generalized quantifiers of type $(1,1)$. Recall from section 2.2 that these are classes of models of a vocabulary with two unary predicates, say A and B. Recall further that for CE-quantifiers, i.e. GQs of type $(1,1)$ that are *conservative* (CONS) and *domain independent* (EXT), the cardinalities $|A \setminus B|$ and $|A \cap B|$ completely determine whether a model $\mathcal{M}$ belongs to that class or not. We captured this by assigning to every such quantifier Q its unique relation R_Q, such that:

$$\mathcal{M} \in Q \Leftrightarrow R_Q(|A \setminus B|, |A \cap B|). \tag{4:1}$$

It was recognized by van Benthem (1986) that we can encode the relevant aspects of such models using strings over an alphabet of two symbols. One symbol encodes elements of $A \cap B$; the other stands for a elements in $A \setminus B$. Concretely, this works as follows (cf. Szymanik, 2009; Steinert-Threlkeld & Icard, 2013; we are still assuming finite models).

Definition 4:2. *Let $\mathcal{M}$ be a model of a vocabulary (A, B), $\vec{a}$ an enumeration of A and $n = |A|$. The function τ maps the pair $(\vec{a}, B)$ to a string $\tau(\vec{a}, B) \in \{0,1\}^n$ such that:*

$$\big(\tau(\vec{a}, B)\big)_i := \begin{cases} 0, & \text{if } a_i \in A \setminus B \\ 1, & \text{if } a_i \in A \cap B. \end{cases}$$

With this encoding in place, every CE-quantifier can be assigned a language $\mathcal{L}_Q$ in the following way.

Definition 4:3. *Let Q be a CE-quantifier, R_Q the corresponding numerical relation and $\#_0$, $\#_1$ functions that map strings to the number of zeros and ones they contain, respectively. We define:*

$$\mathcal{L}_Q := \left\{ s \in \{0,1\}^* : R_Q\big(\#_0(s), \#_1(s)\big) \right\}.$$

A corollary of the equivalence 4:1 (cf. Proposition 2:11) is that Q can be identified with $\mathcal{L}_Q$:

Corrolary 4:4. *Let $Q, \mathcal{M}, A, B, \vec{a}, n$ be as before and $s = \tau(\vec{a}, B)$. Then, $s \in \mathcal{L}_Q \Leftrightarrow \mathcal{M} \in Q$.*

As above in chapter 3, the question whether $\mathcal{M} \in Q$, or in this case also $Q_M(A, B)$, is a standard decision problem and it can be asked what computational resources this problem requires.

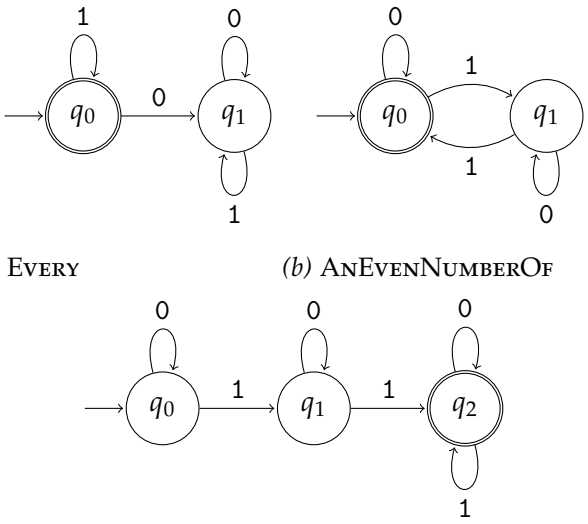
4.1.1.1 *Automata theoretic characterizations of some CE-quantifiers*

To give a simple example, the quantifier EVERY (see section 2.2 for a formal definition) corresponds to the language

$$L_{\text{EVERY}} := \{s : \#_0(s) = 0\}.$$

This is just the language 1^* from section 2.3 and it is recognized by the *deterministic finite state automaton* (DFA) described there and shown again in Figure XVIIa. There are also other automata that recognize L_{EVERY}, but this is the simplest one. Interestingly, all the quantifiers in Aristotle's *square of opposition*, namely EVERY, NO, SOME and NOTEVERY, are recognized by automata that are just as simple. One only needs to relabel the states and transitions. Moreover, these are the only quantifiers that are recognized by such simple devices. All other quantifiers need more states or transitions, at least.

This simple example already indicates that there is regularity; and, in fact, it is possible to characterize the computational require-



(a) EVERY *(b)* ANEVENNUMBEROF

(c) ATLEASTTWO

Figure XVII. The minimal DFA corresponding to some of the discussed quantifiers.

ments of quantifiers in terms of their *definability* (see Definition 2:4). Moreover, there are interesting connections to the *Chomsky hierarchy*. The following theorem identifies definability in first-order logic with recognition by acyclic permutation invariant DFA. This implies that the first-order definable GQs all correspond to *regular languages* (see section 2.3 and Hopcroft & Ullman, 1979 for details concerning the correspondence between classes of languages and automata).

Theorem 4:5 (van Benthem, 1986, pp. 156-157)**.** *A quantifier Q is definable in first-order logic iff L_Q is recognized by a permutation-invariant acyclic DFA.*

Another example of a first-order definable quantifier is ATLEASTTWO (see section 2.2, example 2:2). The simplest automaton that corresponds to this quantifier is shown in Figure XVIIc. The language recognized by this automaton is $\{s \in \{1,0\}^* : \#_1(s) \geq 2\}$. The automaton can be considered an extended version of the one that recognizes SOME, with one extra non-accepting state.

The automaton shown in Figure XVIIb, on the other hand, corresponds to a quantifier that cannot be defined in first order logic. It recognizes the language $\{s \in \{0,1\} : \exists n \in \mathbb{N}(\#_1(s) = 2n)\}$. The automaton is hardly more complex than those in panels a and c of the figure. However, it contains a *non-trivial loop* and such loops are

excluded by the above theorem. In order to be able to define AnEvenNumberOf, we have to enrich first-order logic (see Definition 2:3). This can be done using the *divisibility quantifier* $D_2 := \{(M, A) : \exists n \in \mathbb{N}(|A| = 2n)\}$. More generally, the following theorem holds, which provides a logical characterization of all monadic quantifiers that are recognized by DFA.

Theorem 4:6 (Mostowski, 1998)**.** *DFA accept exactly the class of quantifiers of type* $(1, ..., 1)$ *definable in first-order logic enriched with all* D_n *for* $n \in \mathbb{N}$.

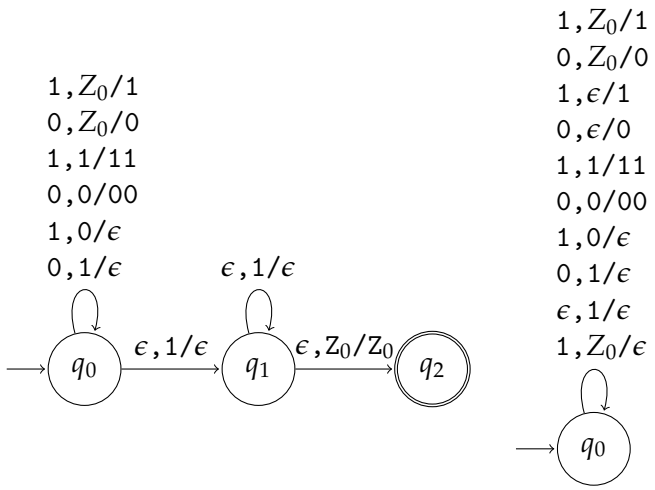
There are also expressions in natural language that correspond to quantifiers not recognized by any DFA. The standard example is the determiner *most* (Barwise & Cooper, 1981) usually assumed to denote the GQ Most (defined in 2:2), which, in turn, corresponds to the non-regular, *context-free* language $\{s \in \{0,1\}^* : \#_1(s) > \#_0(s)\}$. A pushdown automaton (PDA) that recognizes this language was already described in section 2.3 and a slightly different variant of this automaton is shown in Figure XVIIIa. What this automaton does, is to compare two cardinalities. Because it is not known beforehand how large these will be, no finite automaton can perform this task. Instead some kind of memory, like the *stack* of a PDA, is needed. Other similar examples are comparative proportional determiners like *more than half*, *less than a third* or *more than two thirds*. A characterization of the entire class of quantifiers that are recognized by PDA is given below. It is based on definability in the so-called *Presburger arithmetic* (cf. van Benthem, 1986 or Steinert-Threlkeld & Icard, 2013).

Definition 4:7. *A quantifier Q is first-order additively definable if there is a formula* ϕ *in the first-order language with equality and an addition symbol* $+$ *such that* $R_Q(a, b) \Leftrightarrow (\mathbb{N}, +, a, b) \models \phi[a, b]$.

For example, the quantifier AnOddNumberOf, corresponding to the language $\{s \in \{0,1\} : \neg\exists n \in \mathbb{N}(\#_1(s) = 2n)\}$, can be defined using the formula $(\neg\exists x(a = x + x)) \wedge \exists x(x = b)$ and Most can be defined by $(\exists x(a = b + x)) \wedge \neg(a = b)$.

Theorem 4:8 (van Benthem, 1986, pp. 163-165)**.** *A quantifier Q is first-order additively definable iff* L_Q *is recognized by a PDA.*

There is more interesting work on automata theoretic characterizations of natural language quantifiers. For example, Steinert-Threlkeld and Icard (2013) studied iterated quantifiers that correspond to prominent readings of multiply quantified sentences and Kanazawa (2013)



(a) accepts by state

(b) accepts by empty stack

Figure XVIII. Two alternative PDA recognizing MOST.

characterized the monadic quantifiers that are recognized by *deterministic* PDA. But the theoretical distinctions that we focused on so far suffice for the present purpose. These have guided psycholinguistic research on the processing of quantifiers. This is what we turn to next.

4.1.2 *Link to psycholinguistics*

Recently, researchers have started to ask whether the computational distinctions just introduced are also reflected in how natural language quantifiers are processed by humans. Before we summarize some of the relevant experimental work, we briefly consider what links could exist, in principle, between theoretical considerations of the mentioned kind and psycholinguistic studies. In particular, we distinguish three types of links. The first is captured in the following hypothesis.

Hypothesis 4:9. *Computational distinctions between different types of devices needed to recognize natural language quantifiers have psychological relevance.*

This hypothesis is somewhat vague and thus calls for further specification. In order to derive predictions for a concrete test case, it has to be spelled out in more detail. To give an example, it was hypothesized in the literature that quantifiers which cannot be recognized

by DFA but need PDA instead should depend heavily on *working memory* (Baddeley & Hitch, 1974). Thus, enhanced involvement of working memory was predicted to be detectable in experiments (e.g. McMillan et al., 2005, R. Clark & Grossman, 2007 and Zajenkowski, Szymanik, & Garraffa, 2014; see below for more details).

The second potential link is provided by the hypothesis that the discussed automata are – at an appropriate level of abstraction – realistic descriptions of the procedures used by humans to verify or falsify quantified sentences. Obviously, the few examples from above are not sufficient to provide us with a useful hypothesis: their empirical coverage is insufficient by any measure. We need to generalize the approach. One possibility is to specify 'canonical' devices for a wide range of quantifiers. This can be motivated, for example, by optimality consideration along the lines of Anderson's (1989) Rational Analysis (see section 2.1.2; cf. Szymanik, 2016, pp. 51-54). With regard to DFA, this is straightforward: for every regular language a simple recipe (Hopcroft & Ullman, 1979, pp. 67–71) can be used to construct the minimal DFA that recognizes it. With regard to PDA, things are more complicated. Consider the automaton shown in Figure XVIIIa, which recognizes MOST (cf. section 2.3). In addition to this three-state automaton, there is also a one-state automaton that accepts exactly the same strings (shown in panel b of the figure). The one-state automaton has two states less, but the list of stack instructions is somewhat longer. The procedures defined by these two devices may be rather similar. Nevertheless, the canonical procedure has to be specified unambiguously in order to make our second approach workable.

Hypothesis 4:10. *As far as it is specified unambiguously, the 'canonical' automaton is – at some level of abstraction – a realistic description of the procedure used to perform quantifier verification or falsification. In the case of DFA, the minimal automaton is the canonical one. In the case of PDA, the canonical automaton strikes a balance between simplicity with regard to state transitions, on the one hand, and stack operations, on the other.*

As a third alternative, we may associate each natural language quantifier with a small set of automata that recognize it. We may posit that the best performing one is chosen in each instance. For example, Szymanik (2016) noted that a DFA can decide whether $(M, A, B) \in$ MOST as long as elements of A and B are presented in pairs (see also Steinert-Threlkeld, Munneke, & Szymanik, 2015). Under these

special conditions a DFA suffices for the task, although none does in general. This kind of flexibility may be considered an advantage of the third over the second approach. In contrast to the first one, which is only concerned with types of computing devices, the third approach still requires detailed descriptions of the individual verification procedures within the range of possibilities.

Hypothesis 4:11. *Natural language quantifier are associated with small sets of automata that are realistic descriptions of possible procedures used to perform verification or falsification.*

Among these three hypotheses, the weakest claim is made in 4:9, a stronger claim is made in 4:11 and the strongest claim is made in 4:10. In the next section, experimental studies are discussed that tested predictions of the automata model concerning the processing of natural language quantifiers. It may not always be obvious which, if any, of the three approaches was taken in these studies. Nevertheless, the three approaches give the discussion some structure. This is important because, as we will see, the automata model itself is neither unequivocally supported by the experimental data nor can it be refuted. Rather, the interesting question seems to be what an appropriate linking hypothesis is.

4.1.3 *Experimental investigations*

A number of psycholinguistic experiments have tested predictions derived from the automata model. The following two theoretical distinctions play a major role in these studies. Firstly, quantifiers that are recognized by DFA were compared to others that require PDA. Especially the need for a memory device received attention in this regard. The second distinction is that between Aristotelian quantifiers and other more complex ones, e.g. numerical quantifiers. Recall from section 4.1.1 that the quantifiers in the Aristotelian square of opposition exhaust the simplest possible DFA.

4.1.3.1 *PDA and working memory*

As far as I know, the first experiment that tested predictions derived from the automata model was an fMRI study conducted by McMillan et al. (2005). They compared first-order to higher-order definable quantifiers during sentence-picture verification. They reasoned that both types of quantifiers would recruit brain regions in the parietal

cortex that are associated with number knowledge and the processing of numerosity. In addition, higher-order quantifiers were predicted to activate regions in the frontal and prefrontal cortex that are associated with working memory. The idea was, of course, that the theoretically derived discintion between DFA and PDA should be reflected in the mechanisms actually utilized by their participants. Thus, McMillan et al. took the approach of Hypothesis 4:9, stating that the distinctions between different classes of automata are psychologically relevant. This is reflected clearly in the following quote:

> We, thus hypothesize that the qualitative differences between the first-order and higher-order classes of quantifiers, formally reflected in a difference in the computational machinery needed to simulate them, will also be reflected in brain anatomy. (p. 1730)

These predictions were confirmed. In particular, by subtraction of the fMRI signals obtained from the two classes of quantifiers, McMillan et al. found a significant increase of activity in the predicted regions. These results lend support to the linking hypothesis in 4:9. They were supplemented by McMillan, Clark, Moore, and Grossman (2006), who compared the performance of healthy individuals to that of patients diagnosed with a range focal neurodegenerative diseases. The experimental task was similar to that of the first study. The main findings were that (a) patients diagnosed with corticobasal degeneration, a disease that is known to impair number knowledge, performed poorly across both classes of quantifiers (cf. also Troiani, Clark, & Grossman, 2011); and that (b) patients diagnosed with Alzheimer's disease or with frontotemporal dementia, both known to involve working memory deficits, had greater difficulty with higher-order than with first-order quantifiers. However, higher-order quantifiers led to worse performance than first-order definable ones across all groups of participants. Thus, support for the distinct role of working memory in the processing of the latter class is not as clear as in the fMRI study. In addition to these results, the authors also report a positive correlation between a measure of working memory capacity and performance in the higher-order but not the first-order condition across all participant groups. We conclude with the authors that both studies in combination provide experimental evidence for Hypothesis 4:9.

In a comment, Szymanik (2007) highlighted that the distinction between first- and higher-order definable quantifiers is not exactly congruent with the distinction between DFA and PDA (see section 4.1.1, specifically Theorem 4:6). In particular, McMillan et al. (2005, 2006) included parity quantifiers (*an even number of* and *an odd number of*) into their sample of the latter class of quantifiers. These do not require PDA but are recognized by DFA with loops. Differences between higher-order quantifiers that do vs. do not require PDA are not reported in these studies. Therefore, it seems possible that, in comparison to the first-order definable ones, parity quantifiers also require an enhanced amount of working memory. This would disaccord with Hypothesis 4:10, stating that the minimal DFA are used to verify or falsify these quantifiers. It seems equally likely, however, that the reported effects simply would have been larger had the experimental manipulation coincided exactly with the two types of computing devices that are required to solve the task according to the automata model (for further discussion see also Troiani, Peelle, McMillan, Clark, & Grossman, 2009).

Zajenkowski, Styła, and Szymanik (2011) compared a healthy control group to patients diagnosed with schizophrenia. They compared the performance of these two groups in a sentence-picture verification task involving Aristotelian quantifiers, parity quantifiers, modified numerals and proportional quantifiers. Since schizophrenia has been shown to involve working memory impairments (see references in the paper), it was predicted that, especially with proportional quantifiers, the patient group would perform significantly worse, in terms of accuracy and RT, than the control group. The results confirmed these predictions. It is noteworthy in the present context that parity quantifiers did not differ from modified numerals at all but behaved clearly different from the proportional ones. Specifically, the two groups of participants did not differ significantly in accuracy with regard to numerical and parity quantifiers. Moreover, there was no significant difference between these two types of quantifiers. In contrast, proportional quantifiers led to more errors than numerical and parity quantifiers and this effect was reliably larger in the patient group than in the control group. Finally, patients had longer RTs than the control group across all quantifier types, but this effect was extraordinarily large within the proportional quantifiers.

In two further studies, Zajenkowski and Szymanik (2013) and Zajenkowski et al. (2014) correlated several measures of cognitive ca-

pacity with performance in sentence-picture verification. These studies provided a detailed picture of the cognitive capacities that are involved in the verification of different types of quantifiers and, moreover, the results are generally compatible with the automata model. Specifically, they provide more evidence for the distinct role of working memory in the verification of proportional quantifiers. However, as far as I can tell, they go beyond what can be predicted from the model in combination with linking hypotheses of the kind discussed here. In particular, how capacities in short term memory, cognitive control and intelligence affect performance in quantifier verification and what role the type of quantifier plays in this regard does not follow from the above considerations. To predict such associations from the automata model, more sophisticated linking hypotheses are necessary. While the authors discuss some possibilities, these are not worked out in detail.

A finding that is somewhat disturbing was reported by Szymanik and Zajenkowski (2011), who investigated the involvement of working memory in quantifier verification in a dual-task experiment. Participants performed a verification task that involved parity and proportional quantifiers. Before each trial of the verification task, they had to memorize a sequence of either four or six digits that had to be recalled afterwards. As expected, the difficulty of the memory task had a larger effect on performance with proportional than with parity quantifiers. However, performance with proportional quantifiers was better in the condition with six than with four digits. That higher memory load would facilitate the verification task was unexpected. To explain these results, Szymanik and Zajenkowski assume that the six digit condition was so difficult that participants gave up on the memory task, which set free memory resources. If this explanation was correct, one would expect opposite effects in a quasi-replication with lower memory load (with regard to proportional quantifiers Steinert-Threlkeld et al., 2015 show the expected effect of memory load).

Altogether, the mentioned studies provide evidence that quantifiers that require PDA also require an enhanced amount of working memory during sentence-picture verification. All of the above mentioned linking hypotheses would predict this. However, one important issue was glossed over, so far. The difficulty of the proportional quantifiers showed itself in relatively high proportions of errors. The automata model does not predict mistakes nor does it explain differ-

ences in proportions of errors between different quantifiers. Because of this, the data cast doubt on the linking hypotheses 4:10 or 4:11, stating that the automata are realistic descriptions of the verification procedures that are actually used by humans. These hypotheses referred to realistic descriptions at *some* appropriate level of abstraction. And so, they may be amended accordingly in order to align the theory with the data. However, if empirical support to the model is sought in accuracy data, it needs to be spelled out how errors emerge. A first step in this direction, based on *probabilistic automata*, was taken by Dotlačil, Szymanik, and Zajenkowski (2014). In their model, transitions between states are probabilistic events. Estimation of transition probabilities from experimental data provided a good model fit. However, at this point, neither the predictions of this proposal nor its relation to the original automata model are obvious. So, when it comes to proportions of errors, we are left with the linking hypothesis formulated in 4:9, for now. A final point that should be kept in mind is that the proportional quantifiers tested in the mentioned studies were almost exclusively *more than half* or *less than half*. An interesting open question is whether the findings generalize to other types of quantifiers that require PDA or do instead reflect idiosyncrasies of these quantifiers.

4.1.3.2 *Aristotelian vs. other quantifiers recognized by DFA*

The second theoretical distinction of the automata model that has received attention in experimental work is between different kinds of DFA, in particular, between those that recognize Aristotelian quantifiers and more complex ones. Troiani, Peelle, Clark, and Grossman (2009) compared Aristotelian to "numerical quantifiers." The latter class included modified numerals (e.g. *at least three*) as well as parity quantifiers (*an even/odd number of*). In an fMRI experiment, participants evaluated these quantifiers against serially presented visual stimuli, e.g. a series of pictures of balls of different colors. The authors reasoned that the Aristotelian quantifiers can be evaluated based on an "elementary logic system" which consists in a network involving rostral medial prefrontal cortex, responsible for "decision-making about dichotomous events [...] such as attending to [...] exceptional" ones (p. 105), and posterior cingulate cortex, which supports selective visual-spatial attention. With regard to the numerical quantifiers, activation in the intraparietal sulcus, reflecting processing of numerical or magnitude information, and in the dorsolateral

prefrontal cortex, responsible for retention of numerical criteria, was expected. These predictions were borne out. Moreover, activation of the predicted regions was correlated within but not across the two types of quantifiers. Furthermore, these results, which were obtained from healthy individuals, were compared to results obtained from a group of patients diagnosed with corticobasal degeneration. The patients exhibited atrophy of parietal cortex, including the intraparietal sulcus. As predicted, they were selectively impaired on the processing of numerical quantifiers.

In connection with the discussion in the previous section, it is noteworthy that Troiani, Peelle, Clark, and Grossman did not find any differences in neural activity between modified numerals and parity quantifiers. A difference to the fMRI study of McMillan et al. (2005) is that the dorsolateral prefrontal cortex, associated with working memory, was found to subserve the processing of both, modified numerals and parity quantifiers. The authors speculate that the discrepancy may be due to the different modes of presentation of the visual stimuli in the two experiments (i.e. serial vs. parallel).

While Troiani, Peelle, Clark, and Grossman refer to the automata model when motivating their experimental design, it is not obvious what linking hypothesis they assume. The distinction between cyclic and acyclic DFA is disregarded in their research and this seems to be justified by the results. Thus, the authors would presumably not subscribe to linking hypothesis 4:9 in every instance. At the same time, they differentiate between Aristotelian quantifiers, on the one hand, and modified numerals and parity quantifiers, on the other. In terms of minimal automata, this experimental manipulation corresponds to a comparison between the simplest possible DFA, containing just two states, and minimally more complex automata containing additional states or transitions (possibly forming cycles, cf. Figure XVII). This may be taken as endorsement of linking hypothesis 4:10, stating that minimal DFA are realistic descriptions of truth-evaluation procedures. In particular, the description of the "elementary logic system" matches the minimal DFA for Aristotelian quantifiers rather closely, albeit embedded into neural architecture. However, the approach of these authors seems to be more of a mixed one, where predictions are derived from computational considerations in combination with empirical findings. This is reflected in the following quote from Troiani, Peelle, McMillan, et al. (2009):

> [The automata model] thus is not inconsistent with our model, where we emphasize the importance of magnitude processing regions in support of quantifier comprehension. It is clear that both experimental and theoretical approaches can provide complementary evidence regarding the role of quantifier representation in the brain, particularly when computational models are able to furnish concrete predictions about human data. (p. 2685)

Empirically, the distinction between the most simple two state DFA and others seems to have clear consequences. In contrast, the distinction between DFA with and without cycles appears to be less relevant. This is also in accordance with a sentence-picture verification experiment reported by Szymanik and Zajenkowski (2010) (cf. also Szymanik, 2009, 2016), who found that Aristotelian quantifiers were evaluated much faster than parity quantifiers, modified numerals and proportional quantifiers, in the order of mention. The authors conclude from these results that the computational resources (i.e. number of states, loops or memory) needed to recognize a quantifier affect the time needed to evaluate it. Moreover, they suggest that the number of states in the minimal automaton is a more reliable predictor of RT than the requirement of cycles. These conclusion receive additional support from replications reported by Zajenkowski and Szymanik (2013), where modified numerals and parity quantifiers did not differ in terms of RT and accuracy.

4.1.3.3 *Summary and concluding remark*

The experimental studies discussed in the previous two sections show clear processing differences between different types of quantifiers with regard to sentence-picture verification. Proportional quantifiers, which require PDA, are generally more demanding than quantifiers that are recognized by DFA. In particular, we saw evidence for enhanced working memory involvement in the former as compared to the latter class. Moreover, among the quantifiers recognized by DFA, the Aristotelian ones are a special case. They are evaluated relatively fast and are subserved by a distinct network of neural activity. Other distinctions within the group of quantifiers recognized by DFA, e.g. between automata with and without cycles, have less clear effects on processing.

With regard to our linking hypotheses in section 4.1.2, conclusions are somewhat mixed. On the one hand, the fact that the distinction between PDA and DFA was reflected in processing data provides evidence for the linking hypothesis in 4:9, which stated that such distinctions are psychologically relevant. On the other hand, the distinction between automata with and without cycles has only limited reflexes in processing. Thus, the hypothesis has to be restricted accordingly. At the same time, modified numerals differed from Aristotelian quantifiers. Both of these are recognized by DFA. These differences can be accounted for by making reference to their minimal automata. Aristotelian quantifiers correspond to the simplest possible DFA whereas modified numerals need more states and transitions. In this case, a linking hypothesis along the lines of 4:10 or 4:11 seems appropriate, which are both based on canonical verification procedures. However, if we take the minimal automata to be realistic descriptions of the canonical verification procedures, differences in proportions of errors remain to be explained and we have to think about the level of abstraction at which these descriptions are to be located (cf. sections 2.1.1 and 2.1.2).

Before we move on to discuss interface transparency in the next section, I would like to highlight one aspect of the automata model and its relation to empirical studies that was glossed over so far. Recall from section 4.1.1 what decision problem the semantic automata compute. It is the decision problem whether some model belongs to a specific class or not. Finite models are encoded as binary strings, such that each CE-quantifier Q corresponds to a language L_Q over the alphabet $\{0,1\}$. In the experiments, on the other hand, participants saw visual stimuli and had to decide whether sentences are true in that context or not. While it is easily conceivable that the visual stimuli are, in the process of verification, encoded as assumed in the automata model, this (pre-)processing step has received hardly any attention. We tacitly assumed that (*i*) the process of encoding can be neglected; and that (*ii*) the truth evaluation process is based exclusively on the information contained in the assumed encodings.

That these assumptions are not always appropriate was demonstrated by Steinert-Threlkeld et al. (2015). As briefly mentioned above, they noted that the decision problem whether $(M, A, B) \in \textsc{Most}$ can be decided by a DFA if the elements of A and B are encoded as pairs of As and Bs. In a dual-task experiment involving sentence-picture verification of *most* and *more than half* in combination with a digit re-

call task, they showed that working memory involvement is indeed reduced if the objects shown are arranged in pairs. This shows that the encoding of the models (or pictures) can affect experimental results, contrary to assumptions (*i*) and (*ii*). Moreover, it was mentioned above that Troiani, Peelle, Clark, and Grossman (2009) speculated that, with regard to working memory involvement in the verification of modified numerals, differences between their experiment and that of McMillan et al. (2005) may be due to how the visual stimuli were presented. This explanation also takes into account encoding of the stimulus materials. Finally, the confirmed prediction of Troiani, Peelle, Clark, and Grossman (2009) that the verification of Aristotelian quantifiers would involve posterior cingulate cortex because this region supports selective visual-spatial attention also seems to be based on considerations regarding the encoding of the stimuli. There would have been no need for visual-spatial attention, had the stimuli been presented auditively, for example. However, these cases cannot be distinguished in the automata model. In conclusion, it should be kept in mind what the automata model is a model of and what aspects of quantifier verification it does not address in its current form.

4.2 INTERFACE TRANSPARENCY: FROM SEMANTICS TO PSYCHO-PHYSICS

A second line of recent research also investigates the processes underlying the verification and falsification of quantified sentences. The general motivation behind these studies was to use sentence-picture verification experiments as an additional data source that informs us about representations of sentence meaning and thus complements the classical introspective judgments. In particular, these studies aim to differentiate specifications of truth conditions that are logically equivalent – a task that may be hard to come by using introspective judgments. Of course, this is only feasible if there is any connection between the specification of truth conditions and verification procedures at all. Whether this is the case was recently asked by Steinert-Threlkeld et al. (2015), who distinguish the following two alternatives:

> The relationship may be *permissive*: once a specification of truth-conditions is "exported" to general cognition, anything goes. There is no systematic connection between the ways that truth-conditions are specified and judgments of

> truth in context are made. On the other hand, the relationship may be *constrained*: the ways in which truth-conditions are specified correlates with and constrains the methods of verification of sentences in context. (p. 368)

One core hypothesis of the studies discussed in the present section is that quantified sentences are associated with canonical verification procedures which reflect the compositional encoding of their truth conditions. The hypothesis was formulated most explicitly in the *interface transparency thesis* (ITT).

Hypothesis 4:12 (interface transparency thesis, ITT, Lidz et al., 2011)**.** *The verification procedures employed in understanding a declarative sentence are biased towards algorithms that directly compute the relations and operations expressed by the semantic representation of that sentence.*

It will become obvious below that the fundamental idea behind the ITT and related experimental work is that truth conditions alone, as they are usually formulated in semantic theory, may under-determine crucial aspects of sentence meaning. There is thus a connection to the philosophical work hinted at briefly in the introduction, in chapter 1 that identifies sentence meaning with algorithms determining the truth value of a sentence in context (see e.g. Moschovakis, 1994 for discussion).

Experimental studies that provide evidence pertaining to the ITT are discussed here more or less in chronological order. There is some overlap with the discussion in section 4.1, especially regarding questions about canonical verification procedures. Given that the two approaches are concerned with similar questions, it is remarkable that the studies discussed in the present section contain hardly any reference to the automata model. Cases where hypothesized verification procedures correspond closely to one of the semantic automata from above will be highlighted.

4.2.1 *Verification profiles of 'most' and 'more than half'*

As we have seen, GQT provides the means to treat quantifying expressions as atomic (cf. section 2.2). Hackl (2000, 2009) famously argued that by this practice crucial aspects of their meanings are missed. Concerning the quantificational determiner *most*, Hackl (2009) argued that its semantic properties can only be described faithfully if it is

analyzed as the superlative form of *many* (cf. Bresnan, 1973). As compared to the standard GQT analysis, Hackl's semantic analysis assigns a logical form to the sentence in 4:13-a that is relatively complicated. It involves a silent determiner, covert movement and decomposition of *most* into *many* and the superlative morpheme *-est*. However, just as in the GQT approach, the truth conditions of 4:13-a come out equivalent to those of 4:14-a: the two sentences are predicted to be true in the same models.

(4:13) a. Most of the dots are blue.
b. $|Dot \cap Blue| > |Dot \setminus (Dot \cap Blue)|$

(4:14) a. More than half of the dots are blue.
b. $|Dot \cap Blue| > 1/2|Dot|$

Hackl argued that there is nevertheless a subtle, but empirically detectable, difference in how the truth conditions of these two sentences are specified. In particular, these specifications "mimic closely" the relations and operations in 4:13-b and 4:14-b, respectively. Furthermore, he suggested that 4:13-b and 4:14-b correspond to two different "natural algorithms." These are described in the following quote:

> A natural algorithm triggered by [4:13-b] might be a form of vote-counting where subjects simply keep track of whether for each [dot] that is [blue] there is also [a dot] that is not [blue]. If at least one [dot] that is [blue] can be found [without] a counterpart [...] the statement will be true; otherwise it will be false. [4:14-b], on the other hand, might trigger an algorithm that is more akin to checking whether the number of [dots] that are [blue] is bigger than some criterion n which represents half the total number of [dots]. (p. 86)

Of course, the former description corresponds closely to the procedure implemented in the two automata in Figure XVIII. The only potential difference I see is in the type of memory device that is used. This is left unspecified in the quote but defined as a stack in the PDA.

The hypothesis that the verification procedures of *most* and *more than half* differ along these lines was tested using the *self-paced counting* method. Participants first listened to recordings of sentences like 4:13-a and 4:14-a and then iteratively examined groups of colored dots on successive pictures, called "frames." By pressing a button,

they initiated presentation of the next frame. They were instructed to verify or falsify the sentences by pressing one of two response keys. In the critical trials, correct judgments could only be made after the last frame had been examined.

In the first experiment, dots of the two colors were distributed uniformly across frames. While accuracy did not differ significantly between the two quantifiers, RTs on individual frames were consistently and reliably faster for *most* than for *more than half*. Hackl argued that *most* led to faster RTs because the stimulus arrays were particularly well-suited for the hypothesized "vote-counting" procedure. Furthermore, it was predicted that manipulating the distribution of the two colors over the stimulus array would have a larger effect on *most* than on *more than half*. The reasoning behind this prediction was based on the idea that *more than half*, in contrast to *most*, "has a component, namely counting (or estimating) how many half of the dots is, that is constant across all frames, and so should not be affected by the distributional asymmetries at all" (p. 92). This prediction was confirmed in a follow-up experiment that manipulated whether objects of the target color were presented late vs. early in the stimulus arrays.

Hackl drew a two-part conclusion from these experiments. The first part is that, although truth-conditionally equivalent, *most* and *more than half* are associated with different verification procedures. The second is that the verification procedures are indeed as hypothesized. Concerning the first conclusion, it should be noted that there is evidence for differences in meaning between *most* and *more than half* that is independent from the utilized verification procedures (e.g. Ariel, 2004; Kotek, Sudo, Howard, & Hackl, 2011; Solt, 2016a). This conclusion should thus be taken with a grain of salt. Nevertheless, with regard to the ITT, this study was to my knowledge the first to show that the operations and relations in the encoding of truth conditions may constrain verification procedures in predictable ways.

4.2.2 *Verification of 'most' and psychophysics*

Pietroski et al. (2009) also investigated the verification procedures associated with the quantificational determiner *most*. In particular, they asked whether a sentence like 4:13-a (repeated as 4:15-a) introduces a bias towards verification in terms of comparison of set sizes, as in 4:15-b, or in terms of establishing a certain kind of mapping between the blue and non-blue objects, as in 4:15-c.

(4:15) a. Most of the dots are blue.
b. $|Dot \cap Blue| > |Dot \setminus Blue|$
c. $\textsc{OneToOnePlus}_M(Dot \cap Blue, Dot \setminus Blue)$

The relation OneToOnePlus, which I treat here simply as a non-conservative GQ of type $(1,1)$, holds of two sets iff there is a one-to-one correspondence between a proper subset of its first argument and its second argument. As above, the truth conditions in 4:15-b and 4:15-c are equivalent, but Pietroski et al. argue for a natural correspondence to certain verification procedures. In particular, they stress that verification procedures may differ with regard to "representational resources" they involve: The former involves representation of cardinality and the latter involves representations of one-to-one correspondence.

Reasonable candidate procedures for establishing that OneToOnePlus holds between two sets are, again, provided by the semantic automata shown in Figure XVIII. Concerning 4:15-c, on the other hand, Pietroski et al. hypothesized that – at least in an appropriate experimental setting – it would involve the *approximate number system* (ANS, see e.g. Feigenson, Dehaene, & Spelke, 2004). The ANS and its relevant predictions are briefly summarized in the next section, before we discuss the experiment and results of Pietroski et al.

4.2.2.1 *Numerical comparison in cogintive psychology and the ANS*

Cognitive psychology offers a perspective on judgments of numerical inequality that differs from what is implemented in the automata model (if not indicated otherwise, I follow Dehaene, 2007 in my exposition of the ANS). Since the seminal study of Moyer and Landauer (1967) it is assumed that the mental representations of numerical information are noisy and use an analog format. The main motivation for this assumption was to explain so-called *size* and *distance effects* on proportions of errors and RTs in number comparison or discrimination tasks that were observed by Moyer and Landauer and replicated often since:

- *Distance effect*: Distant numbers are easier to compare than closer ones;
- *Size effect*: For constant numerical distance, large numbers are more difficult to compare than smaller ones.

It is often argued that size and distance effects are incompatible with typical digital, computer-like encodings of numbers – or at least unexpected (e.g. Dehaene, 1997, ch. 9).

The hypothesis of noisy and analog mental representations has two prominent formulations. Both share the core assumption that a number or *numerosity* (an alternative term for the cardinality of a set) n is represented by a normally distributed random variable X_n. The *log-Gaussian* model (Dehaene & Changeux, 1993) assumes that the mental representations of numerosities resemble Gaussian probability distributions over a logarithmically compressed *mental number line*.

Hypothesis 4:16 (Log-Gaussian model)**.** *A numerosity n is mentally represented as a random variable X_n which is distributed normally with mean $log(n)$ and variance w^2, for some constant w: $\mathsf{X}_n \sim \mathcal{N}(log(n), w^2)$.*

The variance w^2 does not differ between the different n. In contrast, the *scalar variability model* (Gallistel & Gelman, 1992) assumes that the number line is not compressed, but the variance increases proportionally with n^2.

Hypothesis 4:17 (Scalar-variability model)**.** *A numerosity n is represented as $\mathsf{X}_n \sim \mathcal{N}(n, (nw)^2)$, for some constant w.*

On the basis of these hypotheses, proportions of errors in various tasks that involve comparison or discrimination of numerosities can be predicted using *signal detection theory* (SDT, see section 2.5). Specifically, we may determine the optimal behavior given our assumptions (cf. the Rational Analysis of Anderson, 1989, section 2.1.2). For the case at hand, we consider the task to decide which of two numerosities, n and m, drawn from some set of possible numerosities, is larger. Assuming equal *priors* and a *uniform cost assignment* (cf. section 2.5.1), optimal behavior, in terms of minimizing *Bayes risk* (see Definition 2:30) and thus the probability of error, would be to choose n if the "internal representation" of n is larger than that of m and to choose m otherwise. Consequently, we predict the probability of a correct response to simply be $\Pr(\mathsf{X}_n - \mathsf{X}_m > 0)$, where n is the larger numerosity (cf. Pica, Lemer, Izard, & Dehaene, 2004; Dehaene, 2007, p. 538). Since according to both the log-Gaussian and the scalar variability model X_n and X_m are normally distributed random variables, $\mathsf{X}_n - \mathsf{X}_m$ also has a normal distribution.

As in many other tasks the predictions of the log-Gaussian and scalar variability model are so similar that they are difficult to dis-

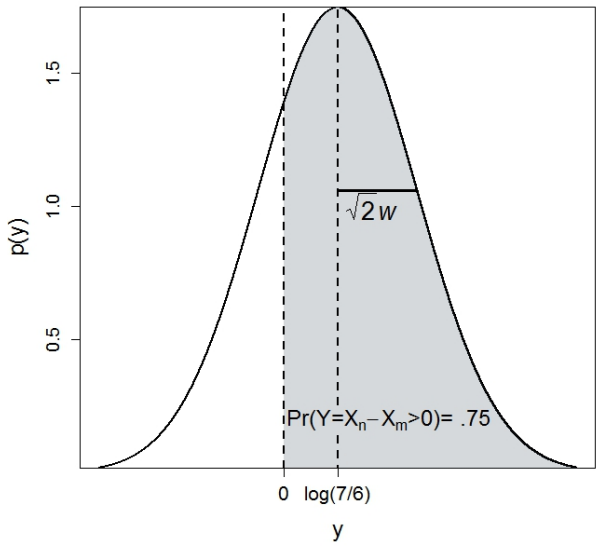
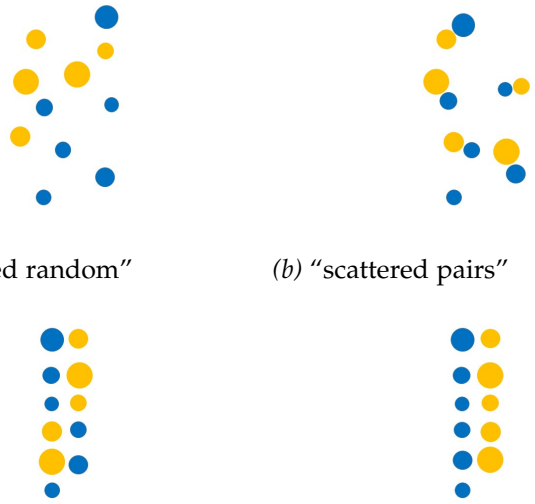


Figure XIX. Sketch of numerosity comparison in the log-Gaussian model at Weber ratio 7:6. The probability density function p of the difference $Y = X_n - X_m$ is shown. The free parameter w is chosen such that $\Pr(X_n - X_m > 0) = .75$, i.e. correct responses are expected in 75% of the cases. It is assumed that X_n and X_m are independent.

tinguish empirically. In both models comparison of numerosities is subject to *Weber's law*, which essentially states that discriminability is ratio dependent. A measure for discriminability is the *Weber ratio*: the ratio at which discrimination reaches some fixed level of performance, e.g. 75% correct discriminations. A Weber ratio that is typical for the comparison of numerosities using the ANS is 7:6 (cf. Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Pica et al., 2004). This case is sketched in Figure XIX. In both models, the Weber ratio, and performance in general, are determined by the free model parameter w.

4.2.2.2 *Comparison of numerosity vs. one-to-one (-plus) correspondence*

Pietroski et al. (2009) conducted a picture verification experiment to test whether there is a bias to verify the quantifier *most* using one of the two verification procedures outlined above. In each trial of the experiment, a picture of yellow and blue dots was shown to the participants for a duration of 150 ms. Their task was to answer the question *are most of the dots yellow* on 360 successive trials. The ratios of yellow to blue dots and their spatial arrangement was manipulated.



(a) "scattered random" *(b)* "scattered pairs"

(c) "column pairs mixed" *(d)* "column pairs sorted"

Figure XX. Types of Visual Stimuli used by Pietroski et al. (2009) with original labels.

There were nine ratios ranging from 2:1 to 10:9 and four different types of arrangements. These four types are illustrated in Figure XXa-d. In the first type, the dots were shown at random positions. In the second, dots were presented in randomly positioned pairs of one yellow and one blue dot in addition to a few remainders of one color. In the third type, two columns of dots were presented and dots of both colors appeared in both columns. Finally, in the fourth type, two columns were presented that were sorted by color. The dependent variable were proportions of correct responses.

It was reasoned that, if participants would tend to answer the question by establishing one-to-one correspondence, the first type of picture (cf. Figure XXa) should be the most difficult one and should lead to the highest amount of errors. From the second through the fourth type of picture (Figure XXb-c) difficulty should decrease. Contrary to this prediction, all conditions except the sorted columns (Figure XXc) led to comparable performance without any significant differences. Furthermore, performance for these conditions was described well ($R^2 \approx .92$) by the psychometric function derived from the scalar variability model of the ANS. The best model fit was achieved for $w \approx .32$. Additionally, the sorted columns led to performance that would be expected for line length comparisons, which involve a smaller parameter value, $w \approx .04$.

The authors concluded that *most* is predominately verified via comparison of cardinality and that this is its canonical verification procedure. Subsequently, Lidz et al. (2011) explicitly formulated the ITT (Hypothesis 4:12) and conducted a follow-up experiment to decide between the specifications of truth conditions in 4:18-b-d.

(4:18) a. Most dots are blue.
b. $|Dot \cap Blue| > |Dot \setminus Blue|$
c. $|Dot \cap Blue| > |Dot| - |Dot \cap Blue|$
d. $|Dot \cap Blue| > \sum_{X \in C} |Dot \cap X|$

Yet again, all these truth conditions are equivalent but involve different operations and thus conceivably may correspond to different verification procedures. All of them involve comparison of cardinalities. But in addition, 4:18-b involves set complementation ($A \setminus B$); 4:18-c involves subtraction of cardinalities ($|A| - |B|$); and 4:18-d, lastly, is based on summation of the cardinalities of the different non-blue color sets of dots (C denotes a *partition* of the non-blue objects according to color).

The experiment was designed to differentiate between a procedure that corresponds to 4:18-d and the other two possibilities. To achieve this, up to three additional colors were shown on the pictures (cf. Figure XXIb). The procedure was identical to that of Pietroski et al. (2009) except that the question was about blue instead of yellow dots this time. It was predicted that performance should decrease as the number of colors increases if the verification procedure would involve estimation of multiple numerosities, as in 4:18-d. This prediction followed from the observation that it is impossible to estimate the numerosities of more than three sets of objects after observing a picture for a time period as short as 150 ms (Halberda, Sires, & Feigenson, 2006). Contrary to this prediction, performance was not affected significantly by the color manipulation. Instead, the scalar variability model explained the data well across conditions ($w \approx .3, R^2 \approx .96$). Furthermore, Lidz et al. showed that a procedure that corresponds to 4:18-c is an especially plausible candidate because it is compatible with previously observed values of w for numerosity comparison. Usually, adults performance in such tasks yields a value of w in the range $0.1 - 0.2$. Lidz et al. explain their larger estimate of w by arguing that the cardinality of the entire restriction set, i.e. the set of dots, enters the computation in the canonical verification procedure, cf. 4:18-c. Since the cardinality of the restriction set is always larger

(a) "2-Colors" *(b)* "5-Colors"

Figure XXI. Types of Visual Stimuli used by Lidz et al. (2011) with original labels. They also used pictures with three and four colors.

than all the individual color sets, *w* is predicted to be larger than if the numerosities of the two color sets would be compared directly. If correct, this implies that the quantifier *most* does indeed introduce a bias to use a suboptimal verification procedure, very much in line with the ITT.

The data of Pietroski et al. (2009) and Lidz et al. (2011) support the ITT. However, one may ask whether it was really the quantfier *most* that introduced the bias to use the sketched verification procedure or whether this bias was, at least to some degree, due to the experimental procedure. In order to address this question, it is crucial to compare different quantifiers. A few studies did this. Firstly, Tomaszewicz (2013) compared different proportional quantifiers in Polish and Bulgarian. Both languages have a proportional quantificational determiner that is equivalent to English *most*, referred to as MOST1. In Polish, this quantifier is written *większość*; in Bulgarian, it is *povečeto*. In addition these languages have a *relative* proportional quantifier which is written *najwięcej* in Polish and *naj-mnogo* in Bulgarian. The latter is referred to as MOST2. Its truth conditions can be described as in the following equivalence, where *C* is again a partition of the non-blue objects according to their color.

(4:19) $\text{MOST2}_M(Dots, Blue) \Leftrightarrow \forall X \in C\big(|Dot \cap Blue| > |Dot \cap X|\big)$

In several experiments, Tomaszewicz investigated the verification of these quantifiers using a similar experimental design as Lidz et al. The finding that is most important for the present discussion was that, as expected, in both languages verification of MOST2 was affected by the color manipulation whereas MOST1 was not. When more than two colors were shown on the pictures, performance dropped signif-

icantly for MOST2 but not for MOST1. These results show clearly that the meaning of a quantifier can affect verification procedures and, furthermore, that verification procedures that depend on cardinalities of multiple color subsets can be executed in these kinds of experiments.

4.2.3 *Back to 'most' vs. 'more than half': "modified interface transperency"*

Kotek, Sudo, and Hackl (2015) studied the verification procedures of the Englisch quantifiers *most* and *more than half* in a sentence-picture verification experiment. The starting point of their experiment was the hypothesis that *most* but not *more than half* allows for a latent, dispreferred relative reading akin to MOST2 in Polish and Bulgarian (Kotek, Sudo, Howard, & Hackl, 2011; Kotek, Sudo, Hackl, & Howard, 2011). They manipulated the ratio of blue to non-blue objects and the number of non-blue colors on a picture. In contrast to the above experiments, there was no time pressure. Although the relative reading was expected to be strongly dispreferred it was reasoned that it should be detectable in the picture verification data. The prediction was that, without time pressure, the color manipulation should affect the processing of *most* but not that of *more than half*. This prediction was confirmed. In the critical conditions, less than half of the objects had the target color. In these conditions *more than half* was overwhelmingly judged as false. With regard to *most*, proportions of judgments depended on the color manipulation. If there were more objects in the target color than in any other color, sentences with *most* were judged as true in a substantial number of trials. This is compatible with the hypothesis that *most* is sometimes interpreted as MOST2.

In order to reconcile their results with the seemingly conflicting results of Lidz et al., Kotek et al. argue that the specific processing demands of an experiment may constrain the readings and verification procedures that are amenable to the participant. They propose a modified version of the ITT in which both the specification of truth conditions and the experimental task may constrain verification procedures (see, however, also Hunter, Lidz, Odi, & Wellwood, 2016 for a very recent response arguing that Kotek et al. (2015) mistakingly conflated the possible truth conditions a sentence can have with the possible verification procedures that can be used to verify these truth conditions):

Hypothesis 4:20 (Modified interface transparency thesis, mITT, Kotek et al., 2015)**.** *When determining the truth or falsity of a statement in a given*

situation, speakers exhibit a bias towards using verification procedures that employ operations specified as part of the truth-conditional import of the statement, as supported by the task demands brought about by that situation.

The final study we discuss in the present section is a sentence-picture verification experiment conducted by Steinert-Threlkeld et al. (2015), who came to a similar conclusion. They compared the verification of *most* and *more than half* against the two types of pictures Pietroski et al. labeled "scattered random" and "scattered pairs" (e.g. XXa-b). As mentioned in section 4.1, Steinert-Threlkeld et al. noted that MOST can be recognized by a DFA if objects are presented in pairs of targets and non-targets, as in Figure XXa. In the general case, DFA are however insufficient to recognize MOST. PDA are needed if objects are presented randomly and no pairing is provided, as in Figure XXb. It was therefore predicted that there should be more working memory involvement in the random than in the paired conditions. This prediction was tested using a dual task experiment: While solving the verification task, participants had to memorize a sequence of digits that was probed at the end of the trial. There were trials with high and with low memory load.

RTs and proportions of errors were lower in the paired than in the random conditions. Moreover, interactions of the memory load and picture type manipulations were found for *more than half* but not for *most*. These results are somewhat difficult to interpret because separate statistical analyses are reported for subsets of the data, but it is not mentioned how multiple comparisons were controlled for. The authors take their findings as evidence for a "constrained relationship between specifications of truth conditions and verification procedures." Moreover, they challenge the claim of Hackl (2009) that *most* is verified in terms of the above mentioned "vote-counting" procedure. If it were, one would expect an interaction between the memory load and picture type manipulations, as was the case for *more than half*.

What may be added to their conclusions is that people can and do make use of visually presented pairings when evaluating *most* without time pressure. This complements the results of Pietroski et al. (2009) who did not find this kind of facilitating effect when visual stimuli were presented for short periods of time. Moreover, the study provides further support for the mITT.

4.2.4 *Summary*

There is accumulating evidence for a constrained relationship between specifications of truth conditions and verification procedures and more specifically for the ITT. This evidence comes in two forms. Firstly, it was repeatedly found that providing additional information that could potentially facilitate or hamper sentence-picture verification (e.g. pairings of objects or additional colors) only shows effects if it is relevant to the purported canonical verification procedure associated with a quantifier (Pietroski et al., 2009; Lidz et al., 2011; Tomaszewicz, 2013). Secondly, it was shown that quantifiers with similar, or even identical, truth conditions may be affected differently by these kinds of manipulations (Hackl, 2009; Kotek et al., 2015; Steinert-Threlkeld et al., 2015). However, there are also some unresolved issues. For example, there is a conflict between the conclusions of Lidz et al. (2011), who do not find an effect of the number of colors in a picture on the verification of *most*, and those of Kotek et al. (2015), who find such effects. Similarly, Pietroski et al. (2009) and Steinert-Threlkeld et al. (2015) report conflicting results with regard to the effect of visually salient pairings on the difficulty of verifying *most*. To resolve such conflicts the mITT was introduced that takes task demands into account beside specifications of truth conditions. A goal for future research is to specify processing models of sentence picture verification in such a way that task demands can be factored in to derive specific predictions.

4.3 CONCLUSIONS

A number of conclusions can be drawn from the preceding discussion. Firstly, processing predictions concerning the verification of quantified sentences against visual contexts can be derived by amending semantic theory with minimal processing assumptions. One approach we have seen was to conceive of quantified sentences and their truth conditions as decision problems and use the theory of formal languages and automata to derive processing predictions. In particular, this approach enables us to predict the minimal computational resources that are needed to solve a verification task under standard semantic assumptions. Another approach was to use probabilistic models from cognitive psychology that are built upon the theory of hypothesis testing and signal detection. In particular, the assumption

that the ANS is involved in quantifier verification was used to predict and model proportions of errors in verification tasks.

Secondly, we have seen that the verification procedures that are actually used are not always the optimal or most efficient ones. Instead, how the truth conditions are specified seems to constrain the range of possible procedures used to verify or falsify quantified sentences. For example, Lidz et al. (2011) concluded from their experimental results that, in order to answer the question whether *most of the dots are blue* a specific verification procedure is used. It involves estimating the cardinality of the non-blue dots. This cardinality is estimated by 'subtracting' the cardinality of the blue ones from the cardinality of the entire restrictor set, namely the dots. This results in worse performance than if the cardinality of the set of non-blue dots was estimated directly. Especially, in case where there are only two colors of dots, this is a striking conclusion because a suboptimal procedure to verify its truth conditions seems to be enforced by the quantifier.

Moreover, there seems to be some variation in what verification procedures are used for a particular quantified sentence. Thus, while the specification of truth conditions introduces a bias to use certain procedures, additional factors seem to be at play that determine which of the possible alternatives is actually used. In particular, task demands may influence which procedures are possible or constrain the available ones further. Two examples we have seen were whether the task introduces time pressure and how the visual contexts are presented.

A number of interesting open questions can be singled out. Concerning the automata model, it is an open question what a suitable linking hypothesis is. We saw that some automata theoretic distinctions are reflected in processing data whereas others are not. Moreover, it is an open question how erroneous performance can be incorporated into this model. Furthermore, it is an interesting question whether and how the ANS and the auomata model can be integrated. For example, what assumption do they share or where do their predictions diverge?

Another open question is how exactly verification procedures of quantified sentences are constrained by truth conditions and task demands in combination. For example: Which are possible verification procedures for a given sentence and which are not? What do the possible procedures have in common? How can task demands favor one over the other? What is needed are explicit models that take

compositional specifications of truth conditions and task demands into account. Such models should identify what aspects of the possible verification procedures are fixed, where variation is possible and what factors affect this variability.

In the following chapter we study a case that poses interesting challenges to the hypotheses and models discussed in the present section. An extension of existing models and a specific implementation of the ITT is proposed that addresses some of the questions just posed.

5

COMPARATIVE MODIFIED NUMERALS: AN EMPIRICALLY MOTIVATED, INTEGRATED PROCESSING MODEL OF TRUTH-EVALUATION[1]

The present chapter is concerned with differences in processing difficulty between UE and DE comparative modified numerals like *more than five* and *fewer than five* in verification tasks (recall the definition of *direction of entailment* from 2.2, Definition 2:8). As we will see, this is an interesting test case to investigate the relation between semantic theory, processing models and experimental data and also one that is particularly informative regarding the questions discussed in the previous chapter. A range of data from psycholinguistic experiments indicate that the DE cases are more difficult to process than the UE ones. While there are some studies that found effects during online comprehension (Bott, Klein, & Schlotterbeck, 2013) or in reasoning tasks (Geurts & van der Slik, 2005), most studies focused on sentence-picture verification or falsification (Koster-Moeller, Varvoutis, & Hackl, 2008; Geurts et al., 2010; Szymanik & Zajenkowski, 2013). These studies consistently found longer RTs for the DE conditions than for the UE ones.[2]

How the increased difficulty of the DE versions comes about is not well understood yet. From the perspective of the semantic automata model (see section 4.1), the increased difficulty of *fewer than n* as compared to *more than n* is completely unexpected. Without introducing *ad hoc* auxiliary assumptions, the automata model does not explain the observed effects. How about ANS models? If the log-Gaussian model of the ANS (see section 4.2.2.1) is combined with models of the

[1] A compact version of the *integrated processing model* and Experiments 3 and 4 were presented at the workshop "Experimental Approaches to Semantics" at ESSLLI 2015 (Schlotterbeck, 2015).

[2] The processing difficulty of other UE and DE quantifiers, beside modified numerals, during sentence-picture verification was also studied experimentally. We restrict ourselves to comparative modified numerals here in order to keep the discussion focused. The discussion is extended to other quantifiers in the following chapter, 6.

processes underlying decisions under uncertainty (namely *sequential sampling models* and most notably the *drift diffusion model* DDM; see Ratcliff, 1978 and section 2.5.2), it accounts well for RTs and proportions of errors in tasks that involve number comparison (e.g. Ratcliff, 2008; Dehaene, 2007). Concerning the processing difficulty of modified numerals, the DDM does, however, only explain the mentioned findings if the model is interpreted somewhat liberally and is, furthermore, enriched with additional assumptions.

In line with the ITT (Hypothesis 4:12), one potential explanation of the mentioned effects would be that they are due to the complexity of the involved semantic representations, or specifications of truth conditions. For example, a range of theoretical proposals assume that DE quantifiers contain covert negation while UE quantifiers do not (e.g. Just & Carpenter, 1971; Jacobs, 1980; Keenan & Stavi, 1986; Penka & Stechow, 2001). Covert negation could lead to enhanced RT because it may correspond to an additional processing step (cf. Just & Carpenter, 1971, Geurts et al., 2010 and also Kaup, Zwaan, & Lüdtke, 2007 for a review of studies on negation in non-quantificational sentences). Similarly, recent proposals in the semantic literature assume an order- or scale-reversing operator in *fewer than* but not in *more than* (Rullmann, 1995; I. Heim, 2006; Büring, 2007a). However, these proposals are generally not formulated as processing models and thus do not make immediate predictions about the processing difficulty of modified numerals. Again, auxiliary assumptions have to be made.

Below, a modified version of the latter proposals is formulated that incorporates minimal and well-motivated processing assumptions. In particular, this *integrated processing model* (IPM) is based on the following hypothesis (cf. R. Clark & Grossman, 2007; Pietroski et al., 2009).

Hypothesis 5:1. *The representation and processing of numerical quantifiers is built upon the representations of approximate numerosity and the processes that operate on them.*

The starting point in devolping the integrated processing model (IPM) are the semantic building blocks usually assumed in theories of comparatives (e.g. von Stechow, 1984; Kennedy, 1997) and comparative quantifiers (e.g. Hackl, 2000). The latter are extended as to incorporate lexical decomposition involving a scale reversing opartor (e.g. Rullmann, 1995). Next, noisy representations of numbers and numerosities are 'plugged in.' This step is similar to what Pietroski

et al. (2009) proposed regarding the quantifier *most*, but the current proposal goes beyond that of Pietroski et al. in two respects. Firstly, it is formulated in a compositional fashion: As a result of introducing noisy representations, other lexical items also receive non-standard interpretations to achieve compatibility. In addition, non-standard computations are used to combine the lexical items. Secondly, verification and falsification processes are described in terms of the sequential sampling models alluded to above (e.g. Ratcliff, 1978). Thus, RTs are modeled in addition to proportions of errors. This is obviously crucial to explain the time required to verify or falsify the modified numerals.

The proposed model also makes novel predictions. These were tested in two experiments. The first (Experiment 3a, section 5.4) collected RTs and proportions of errors in an ordinary sentence-picture verification task. It was designed to disentangle effects of how numerical information is processed from effects due to the direction of entailment – specifically of an additional processing step in the DE conditions. The second (Experiment 4, section 5.5) used the response-signal speed-accuracy tradeoff (SAT) procedure (Dosher, 1979) in order to substantiate conclusions drawn from the first experiment and to decide between alternative explanations. The experimental results provide support for the proposed IPM.

Towards the end of the chapter, the following theoretically relevant points are discussed. Firstly, the essential components of the IPM were motivated independently. This increases credibility relative to other alternatives, at least in the absence of further information. Secondly, because it can be seen as an implementation of the ITT, it provides an interface to classical semantic theory and also to the automata model. Finally, it is embedded in a theoretical framework (see Bogacz et al., 2006, and references therein) that makes reference to all three levels proposed by Marr (1982) (see section 2.1.1) for the analysis of cognitive information processing systems. Thus, it can be extended at different levels.

5.1 PROCCESSING DIFFICULTY OF UE VS. DE COMPARATIVE MODIFIED NUMERALS DURING TRUTH EVALUATION

In several experiments, it was found that sentences with DE comparative modified numerals of the form *fewer than n* take longer to verify or falsify than sentences with their UE counterparts of the form *more*

than n. The first experiment I am aware of was conducted by Koster-Moeller et al. (2008, experiment 1) who used the self-paced counting method (described in section 4.2.1; see also Hackl, 2009) and had participants evaluate sentences like the following.

(5:2) a. More than seven of the dots are blue.
b. Fewer than eight of the dots are blue.

In addition to these, Koster-Moeller et al. also included superlative modified numerals into their experiment, e.g. *at least eight* and *at most seven*. The question they were mainly interested in is closely related to the ITT. They asked whether alternative specifications of equivalent truth conditions (e.g. using *more than seven* vs. *at least eight* or *fewer than eight* vs. *at most seven*) correspond to different verification procedures. From today's perspective, partly due to the data reported by Koster-Moeller et al., the assumption that comparative and superlative modified numerals are semantically equivalent at any level of representation seems questionable (e.g. Geurts & Nouwen, 2007; Geurts et al., 2010; Nouwen, 2010b; Kennedy, 2015), however.

The second question Koster-Moeller et al. investigated was whether and how the direction of entailment influences verification procedures. They derived specific predictions about processing difficulty of UE and DE quantifiers during verification and falsification from a hypothesis that was put forward by Barwise and Cooper (1981, pp. 191–193). We will consider these predictions briefly below, in section 5.1.1. What's important for now is that DE modified numerals were found to be more difficult to evaluate than UE ones, something that does not follow from Barwise and Cooper (1981). In particular, sentences as in 5:2 were evaluated against visual contexts that always contained seven or eight objects in the target color. In each experimental trial, correct decisions could only be made after the last frame was examined. RTs on the final frame were significantly longer for DE *less than eight* than for UE *more than seven*. On average, the DE versions took roughly 200 ms longer to evaluate than the UE ones which led to a significant main effect of the direction of entailment.

In another experiment, Geurts et al. (2010, experiment 3) had their participants verify and falsify sentences like *there are more than two / fewer than three As*. In each experimental trial, a complete sentence was first read self-paced. Afterwards, an array of letters, e.g. As, was examined. Across experimental trials, the number of presented letters was distributed uniformly between the numbers one through

four. The participants' task was to provide a truth-value judgment as fast as possible by pressing one of two response keys. As Koster-Moeller et al., Geurts et al. were also mainly interested in differences between comparative and superlative quantifiers. Again, we focus on the comparative ones here. The direction of entailment did not affect sentence reading times or the proportions of errors, but, crucially, the DE quantifiers did take significantly longer to judge than UE ones. On average, there was a difference of about 250 ms between *fewer than three* and *more than two*. These results are impressive because they are based on only two judgments per condition by 32 participants; and, furthermore, the experimental task does not seem to pose any particular difficulty.

Finally, Szymanik and Zajenkowski (2013) presented Polish sentences that translate as *more than seven / fewer than eight cars are blue* to their participants (see also Szymanik, 2016, pp. 69–74). These were combined with pictures of parking lots showing cars in two different colors. There were always seven or eight target objects, e.g. *blue cars*. [3] Sixty-nine participants provided four judgments per condition each. Again, the DE and the UE conditions did not differ in sentence reading times or proportions of errors, but the DE conditions did take significantly (about 750 ms) longer to judge than the UE conditions.

5.1.1 *A note on interactions with truth values*

In the present chapter, the main focus is on the enhanced difficulty of *fewer than n* as compared to *more than n*. In addition to this main effect, both Koster-Moeller et al. (2008) and Szymanik and Zajenkowski (2013) also found that the direction of entailment interacted with truth-values (superficially similar to effects found in studies on negation, cf. Kaup et al., 2007). However, the form of the interaction differed between the two studies. A short comment is in order here on how to reconcile these seemingly incompatible findings (cf. also Szymanik, 2016). This short digression also provides additional motivation for the reflection on potential processing models in the next section.

Koster-Moeller et al. found that within the conditions that required a "yes, true" response, UE quantifiers were judged faster than the DE ones, but in the false conditions the opposite pattern was

[3] I use the term *target object* to refer to objects that have both the property expressed in the restriction and the scope of a modified numeral.

observed. These results had been predicted based on Barwise and Cooper (1981). In contrast, Szymanik and Zajenkowski found that, within the false conditions, the DE modified numerals were judged considerably slower than UE ones whereas RTs did not differ within the true conditions.

In accordance with the mITT (Kotek et al., 2015, Hypothesis 4:20), these discrepancies are plausibly due to divergent task demands of the self-paced counting and the ordinary picture verification task (cf. Szymanik, 2016, pp. 74–76 for an explanation in terms of task demands within the automata theoretic framework). In particular, self-paced counting presumably involves relatively high memory load (cf. the discussion by Troiani, Peelle, Clark, & Grossman, 2009 concerning different modes of presenting visual stimuli that was referred to in section 4.1.3). Moreover, repeated motor responses have to be prepared and executed taking into account anticipation of yet unseen visual information. Importantly, these factors may affect the UE and DE quantifiers differently in the true and false conditions. For example, compare the case in which *more than seven* is true to that where *fewer than eight* is true. In the latter, extra time may be spent before the truth-value judgment is given in a self-paced counting trial in order to ensure that all target objects have indeed been taken into account. If some are not remembered or have not yet been uncovered, an erroneous judgment would be the result. In the false cases, the situation is exactly reversed. In ordinary sentence-picture verification, such effects are not expected or should at least be much smaller because target objects are easily accessible. On the other hand, RTs in ordinary sentence-picture verification reflect at least two processing components: the process of deciding between "yes, true" and "no, false" and the process of cardinality estimation or counting. The latter is most likely responsible for the interaction effect observed by Szymanik and Zajenkowski. The self-paced counting method allows for separation of these two processing components.

What's crucial to recognize here is the importance of explicit processing models. In the absence of an explicit model that takes into account the specific demands of the experimental tasks, the mentioned findings may appear incompatible. But a completely natural explanation may be found once such a model is specified. Obviously, explicit models are for similar reasons also a prerequisite to sound interpretations of the experimental data with regard to the hypotheses one aims to investigate .

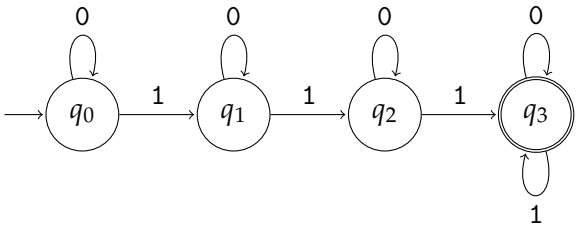
5.2 CANDIDATE PROCESSING MODELS

The present section discusses potential processing models of the truth evaluation of modified numerals. It is asked whether and how these models can explain the experimental results discussed in the previous section. Furthermore, ways of amending existing models to accommodate the experimental result are highlighted.

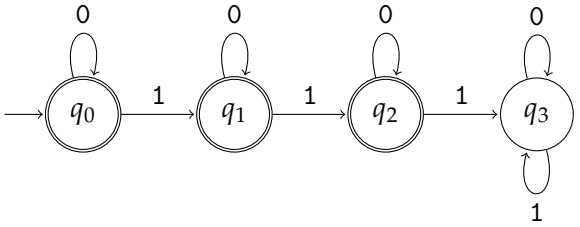
5.2.1 *Automata model*

Under the automata model, the mentioned results are surprising. Figure XXII depicts and explains the minimal DFA for the quantifiers *more than two* and *fewer than three* (see also section 4.1). They execute certain counting procedures. The minimal automata for other modified numerals work completely analogous. Under none of the linking hypotheses discussed in section 4.1.2, namely hypotheses 4:9 through 4:11, are any differences in RTs expected between the DE and the UE modified numerals. If we adopt the strongest of the three linking hypotheses, i.e. 4:10, which states that the minimal DFA are realistic descriptions of verification procedures, the time needed for verification or falsification is predicted by the number of processing steps the automata take. Obviously, the automata for the DE and UE modified numerals need the same number of steps. How many are needed only depends on the number of presented objects. As this was counterbalanced across conditions in the mentioned experiments, the prolonged processing time of the DE conditions is incompatible with this linking hypothesis. Moreover, since the minimal DFA implement the simplest possible verification procedures for modified numerals under their standard semantics, another conclusion can be drawn: Either the assumed semantics of the modified numerals are not accurate or they are not verified or falsified using the simplest possible procedures. This may be considered indirect evidence for the ITT: How truth conditions are specified determines how they are verified or falsified.

To explain the enhanced difficulty of DE vs. UE modified numerals, Szymanik (2016) suggested the following *ad hoc* hypothesis: "'passing through accepting states' is more difficult than 'passing through rejecting states"' (p. 73). This hypothesis predicts that the effect of the direction of entailment should increase with the numeral because larger numerals require more states. Whether this prediction



(a) MoreThanTwo



(b) FewerThanThree

Figure XXII. Semantic automata for the modified numerals *more than two* (top) and *fewer than three* (bottom). To explain how they work: If there is one target object and no other objects, both automata would read `1` and move to q_1. Thus, after two steps, the upper automaton would return `false` and the lower one would return `true`. If there are three target objects and no other restrictor elements, the computation would take six steps. The upper automaton would return `true` and the lower one would return `false`.

is correct remains an open question. Apart from this empirical question, the explanation also faces a conceptual problem. If we combine multiple DE quantifiers, difficulty seems to increase. For example, compare 5:3-a to 5:3-b. Intuitively, the latter sentence seems to be more difficult to understand than the former, but they are logically equivalent.

(5:3) a. Every boy tickled some girl.
b. No boy tickled no girl.

Assuming the observed intuitive difference in quantificational complexity is empirically valid (cf. Bott et al., 2013; Bott, Schlotterbeck, & Klein, n.d.), how can this difference be accounted for in the automata model? Even if the model is augmented with the just mentioned hypothesis, this seems non-trivial because the simplest DFA for the two sentences are identical (see Steinert-Threlkeld & Icard, 2013, p. 169).

5.2.2 *Sequential sampling models of number comparison*

As was described in section 4.2.2.1, cognitive psychology offers models on the processing of numerical information that allow us to predict performance in tasks that involve comparison of numerosities. In particular, it is assumed in these models that the ANS is based on analog and noisy representations of numerosity (see Moyer & Landauer, 1967). Based on this assumption, we can use SDT (see section 2.5) to determine optimal behavior in numerosity comparison tasks and thus derive predictions about performance (e.g. Dehaene, 2007). Moreover, we saw in section 4.2.2 that performance in verification of certain natural language quantifiers can also be modeled successfully this way (Pietroski et al., 2009; Lidz et al., 2011; Tomaszewicz, 2013).

For the case at hand, we aim to describe RTs in addition to proportions of errors. Interestingly, similar considerations that allow us to derive predictions about proportions of errors in terms of SDT, namely what would be the optimal response strategy given our assumptions, can also be used to derive a processing model that predicts RT. One assumption that is added to this end is that decisions are based on a sequence of observations sampled repeatedly from a noisy representation. This naturally leads to a description of the decision process as a *sequential probability ratio test* (SPRT) (see section 2.5.2 and also Bogacz et al., 2006). The general idea behind using the SPRT as a model of decision making is that people start out uncertain about their decision – although there may be response biases – and then gradually move towards one of the response alternatives while they accumulate a noisy signal during the decision process. The rate of accumulation (or drift rate or step size) is determined by the *log likelihood ratio* (LLR, see section 2.5.1) of the two response alternatives – or, more accurately, the two hypotheses – given the current observation.

The decision is made as soon as the LLR crosses one of two predefined response boundaries, i.e. a certain level of confidence is reached. A famous implementation is the *drift diffusion model* (DDM, also explained and motivated briefly in section 2.5.2) of Ratcliff (1978), which can be considered the continuous time analog of the SPRT. Because the DDM implements the optimal response strategy, it can be considered a parsimonious and well-motivated processing model (cf. the Rational Analysis of Anderson, 1989). Thus, it may not come as a surprise that the processes described by the model, in particular, ac-

cumulation of a noisy signal to form a decision variable, were identified in neural circuits involved in perceptual decision making (see e.g. Gold & Shadlen, 2001 and also Forstmann et al., 2016 for a review).

In the DDM, the total response time is the sum of a *non-decision time* including perception, motor processes, *etc.* and a *decision time*, the time needed by the diffusion process for response selection. The diffusion process is governed by the *drift rate*, the *response bias* (or relative starting point) and the *decision boundaries*, which specifies how conservative subjects are in providing a response. The DDM was successfully used to model various aspects of decision processes in numerosity comparison tasks (for discussion see Dehaene, 2007, pp. 543–551 and references there). In particular, *size* and *distance effects* (see section 4.2.2.1) on proportions of errors and on RTs can be explained jointly. Under the assumption of log-Gaussian representations, one general prediction concerning comparison of numerosities is that the drift rate should be proportional to the log-ratio of the two numerosities (Dehaene, 2007, pp. 548).

5.2.2.1 *Application to modified numerals*

How could the verification or falsification of modified numerals proceed given these assumptions? We may think of the evaluation of *more than n* against a context with m target objects as follows. The semantic interpretation of *more than n* introduces a numerical criterion of $n + 0.5$ to which m is compared. The mean step size or drift rate is then predicted to be proportional to $\log((n + 0.5)/m)$. [4] It is positive if there are more than n target objects and negative otherwise. A positive drift approaches the decision boundary for the"yes"-response while a negative drift approaches the "no"-response. The evaluation of *fewer than n* against m target objects could proceed similarly. Here, the decision criterion would be $n - 0.5$ and the mean step size would be $-log((n - 0.5)/m)$.

Applying this model to the discussed experimental data on comparative modified numerals (Koster-Moeller et al., 2008; Geurts et al., 2010; Szymanik & Zajenkowski, 2013), we first note that the experi-

[4] Note that a drift rate of $log(n/m)$, which would be predicted if the task was to choose the larger of two numerosities, would not do the job. On the assumption of this drift rate, participants would be purely guessing when they evaluate *more than n* against n target objects. This is clearly inadequate. Note, however that adding (or subtracting) a constant value of 0.5 to (or from) n is non-trivial when we operate on the logarithmic scale (Dehaene, 2007, pp. 561ff.). This is problematic if we want to derive the criteria $n \pm 0.5$ compositionally from *more than* and n. An alternative model that overcomes this problem is discussed below in section 5.3).

ments share a similar, symmetrical design. In all three experiments, quantifiers of the form *more than n* and *fewer than n+1* (e.g. *more than seven* and *fewer than eight*) had to be evaluated. Both introduce the same numerical criterion $c = n + 0.5$. The two quantifiers had to be evaluated against pictures showing $c \pm a$ target objects (e.g. seven and eight target objects). The corresponding drift rates are predicted to be proportional to $\pm log(c/(c+a))$ and $\pm log(c/(c-a))$. Since *ceteris paribus* drift rates with larger absolute values produce faster decisions, we see that *more than n* should be judged faster when it is false (i.e. evaluated against $c - a$ target objects) than when it is true (i.e. evaluated against $c + a$ target objects) and that the opposite holds for *fewer than n*. The overall decision times for the UE and DE conditions should be symmetrical and no effect of the direction of entailment is expected.

However, if we additionally take into account the uncontroversial assumption that there is a general bias towards "yes"-responses in verification tasks, the effect of the direction of entailment is accounted for. In the DDM, response biases are modeled as diffusion processes that start closer to one of the response boundaries than to the other. A bias towards "yes"-responses implies that these are generally faster. In consequence, the difference in drift rates between the two false conditions has a larger effect than the difference between the two true cases. In line with the experimental findings, the cases with *more than n* are thus judged faster on average than the cases with *fewer than n*. For *fewer than n*, the slower drift rate coincides with the more difficult "no"-response which amplifies the relative difficulty of this condition. In contrast, the relative difficulty of the true *more than n* condition is not pronounced as much because, as an effect of the bias, the boundary for "yes"-response is reached relatively fast.

A POTENTIAL OBJECTION. At this point, the reader may object that the model is implausible. In particular, it relies on the ANS whereas the *exact number system* (see e.g. Feigenson et al., 2004) seems to be at play in the reported experiments: In all three experiments, the number of presented target objects was close to the decision criterion and still there were only few errors. Furthermore, in the experiment of Koster-Moeller et al. (2008), exact counting was invited by the self-paced counting task; in the study of Geurts et al. (2010), the tested numerals were so small that exact representations should be provided by the *subitizing* system (Kaufman, Lord, Reese, & Volkmann, 1949;

Feigenson et al., 2004); and, finally, the data of Szymanik and Zajenkowski (2013) may seem incompatible with approximation since RTs were relatively long.

While this is a reasonable objection, we should bare in mind that the distinction between exact and approximate representations is not always clear-cut. With regard to subitizing, there have been lively debates as to what its status is (e.g Balakrishnan & Ashby, 1992; Revkin, Piazza, Izard, Cohen, & Dehaene, 2008). Furthermore, the original experiment of Moyer and Landauer (1967), which established the empirical foundation for noisy representations of numbers in the first place, employed a task where Arabic numerals had to be compared. Contrary to fact, this may often be expected to trigger exact rather than approximate representations. Finally, the DDM can – although not the norm – produce long decision times with arbitrarily few errors if the decision boundaries are spread apart sufficiently wide.

Putting these issues aside for now, the least we can say is the following. Given the just outlined model of how modified numerals are evaluated, we should be able to design experiments almost identical to the above-mentioned which produce apparent effects of the direction of entailment that are in fact due to number processing. This should be possible just by forcing participants to rely on approximate representations. For illustration, Figure XXIII sketches two sample paths of the diffusion processes and the expected distributions of RTs for the four conditions in such a hypothetical experiment. These considerations suffice to motivate an experimental design that can disentangle effects of the semantic complexity of modified numerals from effects that are caused solely by the processing of noisy numerical information (see Experiment 3a, section 5.4).

5.2.3 *Application of the ITT: additional operators*

Adopting the ITT, a potential explanation could be based on additional semantic operations that take part in the specification of truth conditions of sentences containing *fewer than* but are not involved in *more-than* constructions. Two examples of such operations, namely covert negation and antonym operators, are discussed here.

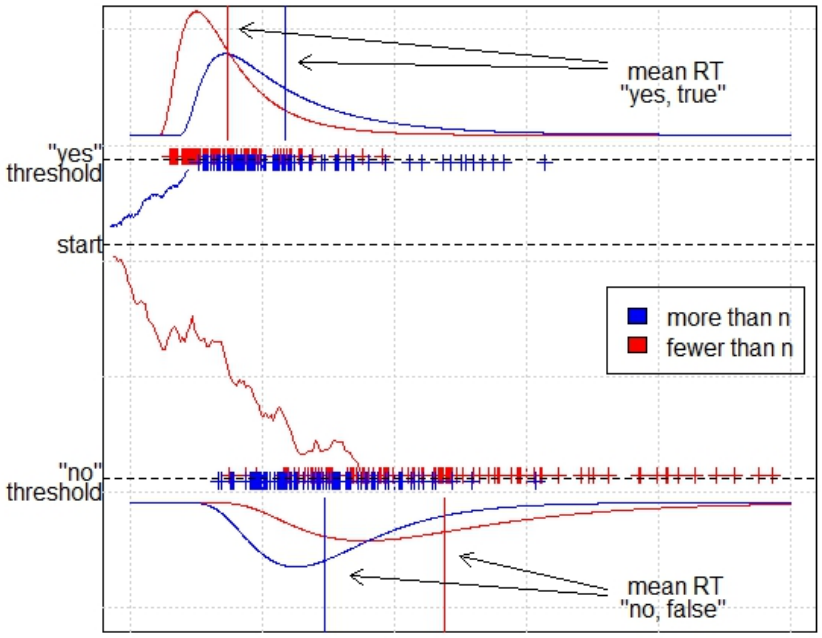


Figure XXIII. In the center of the figure two sample paths of a diffusion process are sketched that correspond to the evaluation of comparative modified numerals in a hypothetical experiment. At the top blue and red crosses show first passage times sample paths crossing the "yes"-boundary when *more than n* (blue) or *fewer than n+1* (red) is evaluated. The blue (*more than n*) and red (*fewer than n+1*) curves show the corresponding expected distribution of RTs for the "yes"-responses. The vertical lines show the means. At the bottom the same is shown for "no"-responses (blue: *more than n*, red: *fewer than n+1*). The data were simulated using the `Rwiener` package (Wabersich, 2014) for R (R Core Team, 2015). In addition to what is described in the running text, it was assumed that a larger number of to-be-estimated objects leads to longer non-decision times (Dehaene, 1997).

5.2.3.1 *Covert Negation*

Several theoretical proposals argue for covert negation in DE quantifiers for a variety of reasons (e.g. H. Clark & Chase, 1972; Jacobs, 1980; Keenan & Stavi, 1986; Penka & Stechow, 2001). One famous

example is the assumption that negative indefinites (e.g. German *kein*) are composed of negation and an existential quantifier. Covert negation may, in turn, correspond to an extra processing step in the canonical verification procedure associated with a quantifier. That such an additional step would cause observable difficulty is plausible since numerous studies have found enhanced processing difficulty in negated sentences (see e.g. the review by Kaup et al., 2007 and also Just & Carpenter, 1971; H. Clark & Chase, 1972). In fact, this is how Geurts et al. (2010) interpret their data.

Processing models that attribute enhanced difficulty of DE quantifiers to covert negation may take various forms. For example, both the automata model and the sequential sampling model described above can be amended with an extra processing step that switches "yes, true" to "no, false" and *vice versa* before the final truth-value judgment. Generally, any processing model of quantifier verification can be amended this way because (*i*), per definition, negation swaps truth values, and (*ii*) any DE quantifier is equivalent to the negation of a suitable UE counterpart (see e.g. Peters & Westerståhl, 2006, ch. 5). For example, in terms of GQT, FewerThanThree is logically equivalent to ¬MoreThanTwo.

Especially if they are based on the ITT, processing models of quantification should be plausible with respect to classical semantic considerations. It is noteworthy that Penka (2011, 2012), who endorses decomposition of the German Aristotelian quantifier *kein* (*no*), considers decomposition of *fewer than n* into *not more than n-1* implausible. Among other things, her reasoning is based on the fact that this kind of decomposition is morphologically opaque. Furthermore, decomposition of *fewer than n* into *not at least n* is implausible on empirical grounds: It has been shown in reasoning and verification tasks that comparative and superlative modified numerals like for example *more than five* and *at least six* are neither semantically equivalent nor interdefinable (e.g. Geurts et al., 2010).

5.2.3.2 *Antonym operators*

Recent semantic theory posits an operotar called FEW or LITTLE, respectively, in constructions containing *fewer than* or *less than*. (e.g. Rullmann, 1995; I. Heim, 2006, 2008; Büring, 2007a, 2007b; Solt, 2009, 2014; Beck, 2012b; Penka, 2015). When applied to *gradable adjectives* (see section 2.6), these operators yield the respective anotnyms. Thus, they are often referred to as *antonym operators* or *antonymizers*. They

constitute an alternative to covert negation that is arguably more plausible from a linguistic stand point. [5] In the present section, a short summary is given of how antonym operators are motivated with regard to ordinary comparatives, which involve gradable adjectives (see section 2.6). After that, I sketch the application of the syntax and semantics of ordinary comparatives to comparative modified numerals, as suggested, for example, by Hackl (2000). Below, in section 5.3, a processing model is developed that contains a processing step corresponding to FEW.

LITTLE IN COMPARATIVES. In her seminal syntactic analysis of comparative constructions, Bresnan (1973) proposed that *fewer* is the phonological spell-out of *-er few* and *less* that of *-er little*. The morpheme *-er* is called the *comparative morpheme*. Thus, *-er little*, pronounced as *less*, is the comparative form of *little*; and *-er few*, pronounced as *fewer*, is the comparative form of *few*. On the basis of this assumption, Bresnan postulated identical morphosyntactic structure for *less* as for expressions like *very little*, *so little* or *as little*. Likewise, *fewer* is completely parallel to *very few*, *so few* or *as few*. Today, Bresnan's assumptions are widely adopted. The following discussion is restricted to *less*, but *fewer* works just the same in most respects (for discussion of the differences between these two lexical items see e.g. Solt, 2009). Moreover, the discussion is based on English, but it applies equally to German, which was tested in the experiments reported below.

Rullmann (1995) suggested that, once we acknowledge that *less* spells out *-er little*, ambiguities in sentences like 5:4 are naturally accounted for. He observed that sentences of this kind may receive one of the two readings exemplified in 5:4-a and 5:4-b whereas the ambiguity disappears if *less high* is replaced with *higher* (see also Seuren, 1979; a collection of natural examples was provided by Büring, 2007b; for a questionaire study using German materials see Beck, 2012b).

(5:4) The helicopter was flying less high than a plane can fly.

a. The height of the helicopter was below the maximum height planes can reach.

b. The height of the helicopter was below the minimum height planes can reach.

[5] Usually, no such operator is assumed in *more-than* constructions although in some proposals a semantically vacuous identity function, MUCH or MANY, is assumed to be part of the semantics of *more than* (Büring, 2007a; Solt, 2014).

Rullmann explains this ambiguity as follows. Phonologically, *less high* is unambiguous and spells out *-er little high*. Semantically, however, *little* has two compositional options. Firstly, it can compose with the comparative morpheme *-er* to its left. Alternatively, it can compose with the gradable adjective *high* to its right. The former option leads to the reading in 5:4-a whereas the latter leads to 5:4-b. Thus, the ambiguity in 5:4 lends support to the decomposition hypothesis.

To see how Rullmann's explanation works concretely, we take a look at the syntax and semantics of comparative sentences (see also section 2.6). One common assumption is that sentences like 5:4 are subject to *comparative deletion* (Bresnan, 1973), a generative rule that elides material from the *comparative clause*, i.e. the clause spanning the words from the immediate right of *than* until the end of the sentence. The contrast between 5:5-a and 5:5-b illustrates the effect of this rule. A key idea in Rullmann's account is that there may sometimes be multiple alternatives as to how the elided material can be recovered when interpreting sentences that are subject to comparative deletion.

(5:5) a. Elin's hands are wider than Simon's feet are long.
b. Elin's hands are wider than Simon's feet are (*wide).
(cf. Kennedy, 2002)

Another common assumption is that comparative clauses contain a silent operator related to a gap inside the clause. Specifically, Rullmann adopts an influential suggestion by Chomsky (1977) that an instance of *wh-movement* is observed in comparative clauses. Chomsky's suggestion is supported by the observation that comparative clauses are subject to similar constraints as constructions involving *wh*-movement, for example so-called *island constraints*. For brief illustration, compare the *wh-island* in 5:6-a to the non-island complement clause in 5:6-b (for further discussion see also Kennedy, 2002).

(5:6) a. *The bag was heavier than I asked whether it was.
b. The bag was heavier than I claimed it was.

These assumptions are illustrated in Figure XXIV, which shows two alternative structural representations that mediate between the sentence in 5:5 and its possible semantic interpretations. Crucial to Rullmann's explanation of the ambiguity is the following. Depending, on whether *little* is combined with the comparative morpheme *-er*, as shown in Figure XXIVa, or with the gradable adjective *high*,

as shown in Figure XXIVb, the elided material is recovered as *high* or *little high*, respectively. These two possibilities correspond to the two possible readings of 5:4. Following a suggestion of von Stechow (1984), OP is semantically interpreted as a maximality operator (or *supremum*, if needed) and the comparative clause is thereby interpreted as a maximal *degree* (which one can think of as real numbers, see section 2.6.1 for discussion). Furthermore, *little* is an order-reversing operator, which can be implemented as a mapping from degrees, d, to their inverse (with respect to addition), $-d$ (see I. Heim, 2006). Skipping some detail for now, the two possibilities in 5:6-a,b thereby encode the truth conditions of the two readings:

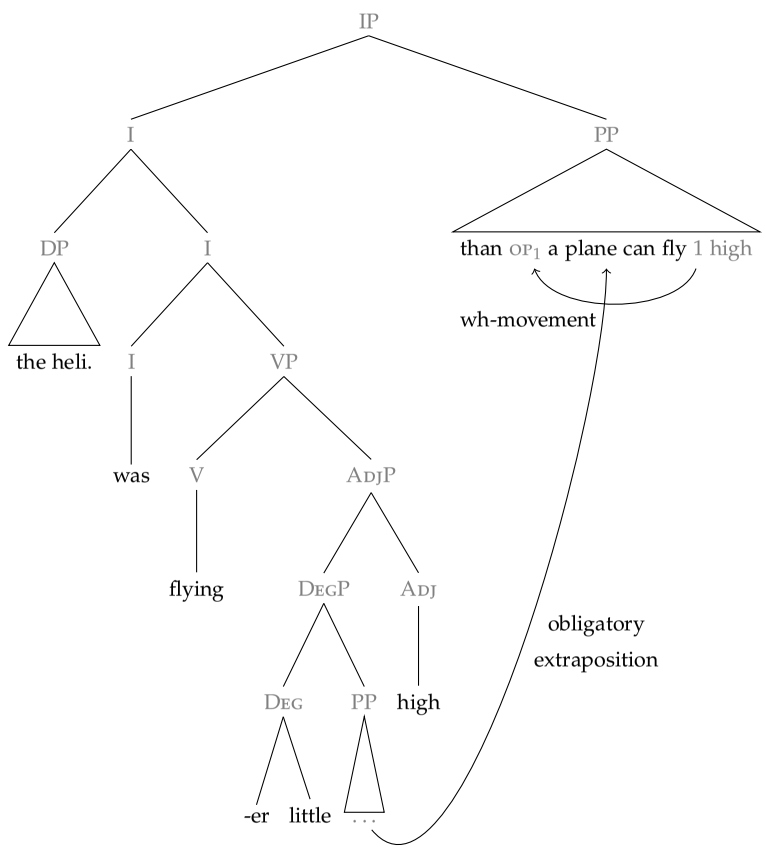
(5:7) a. $\exists d\big(height(heli) = -d \wedge d > -max\{d : a\ plane\ can\ fly\ d\ high\}\big)$
b. $\exists d\big(height(heli) = -d \wedge d > max\{d : a\ plane\ can\ fly\ -d\ high\}\big)$,

which are equivalent to the somewhat simpler:

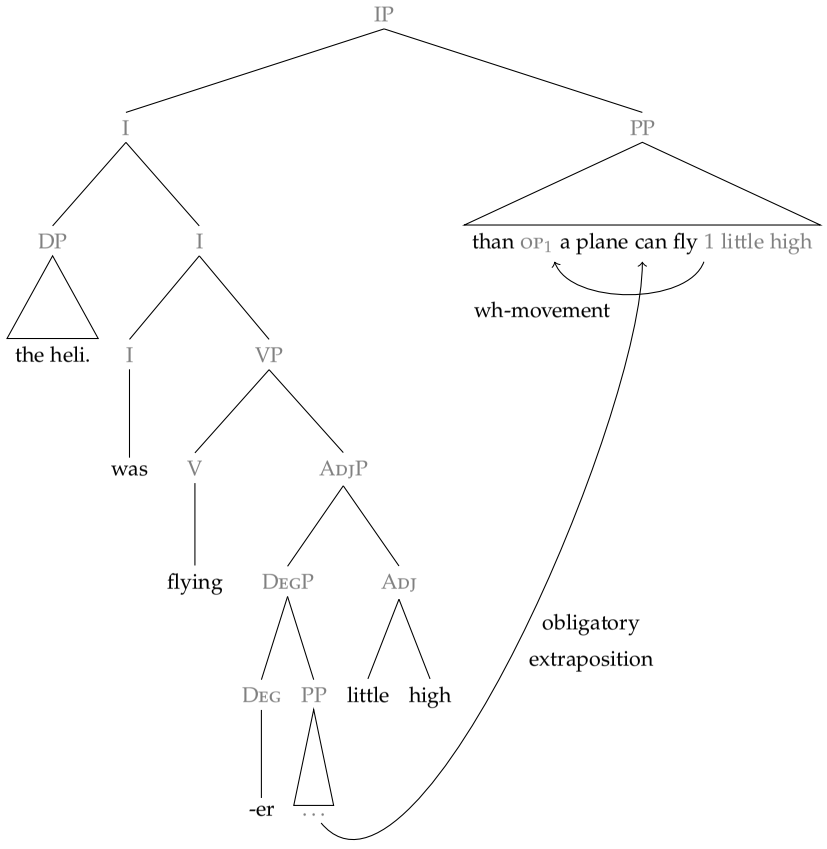
(5:8) a. $height(heli) < max\{d : a\ plane\ can\ fly\ d\ high\}$
b. $height(heli) < min\{d : a\ plane\ can\ fly\ d\ high\}$.

While agreeing with Rullmann's (1995) basic idea, I. Heim (2006) criticized his account as being non-compositional. Alternative, compositional accounts in the same vein have been suggested (I. Heim, 2006, 2008; Büring, 2007a; Solt, 2014) with a slightly different lexical semantics – referred to as "degree negation" – for *little*. I do not share I. Heim's (2006) concern that Rullmann's account is "fundamentally, and not just superficially, non-compositional" (p. 46). For example, Büring's (2007a) proposal can easily be adapted to Rullmann's original assumptions. In appendix A the ambiguity is derived compositionally from the two structures shown in Figure XXIV using both Büring's (2007a) proposal and an adapted one that incorporates Rullmann's assumptions.

Further arguments in favor of decomposition using *little* are also given in the papers just referred to. Among them, Büring's argument concerning *cross polar nomalies* may be the most impressive one. He has shown that decomposition of antonyms using *little* accounts naturally for surprising variation in the semantic well-formedness of comparatives that involve adjectives of opposite polarity, as in 5:9. The unacceptability of 5:9-b has been attributed to *incommensurability* of such cross-polar adjectives (e.g. Kennedy, 2001). This leads one to expect 5:9-d to also be inaccpetable. Büring suggested that,



(a) Maximum reading



(b) Minimum reading

Figure XXIV. Anatomy of an explanation of the Seuren-Rullmann ambiguity: The word *less* is decomposed into *-er little*; OP is the mentioned silent operator, which stands in dependency relation to the co-indexed gap.

in 5:9-d, *shorter* spells out the same type of lexical sequence, *-er little long*, as in the Seuren-Rullmann ambiguity above. Again, *little* may compose with the compartive morpheme *-er* to its left or with the gradable adjective *long* to its right. Thus, in contrast to 5:9-b, 5:9-d has a well-formed reading (the one where *little* composes with *er*) despite involving comparison of cross-polar adjectives at the surface.

(5:9) a. The rope is longer than the gap is wide.
b. #The rope is longer than the gap is narrow.
c. The rope is shorter than the gap is narrow.
d. The rope is shorter than the gap is wide.
(from I. Heim, 2008)

APPLICATION TO COMPARATIVE MODIFIED NUMERALS. How can these assumptions about comparative constructions be applied to the comparative modified numerals *more than n* and *fewer than n*? Based on surface similarity, selectional restrictions and also split-scope data, Hackl (2000, 2002) argued that these determiners are best analyzed using the syntax and semantics of ordinary comparative constructions – a plausible and today widely accepted view (excellent overviews of this and related issues were given by Nouwen, 2010b and Szabolcsi, 2010). Specifically, Hackl suggested that sentences like the following have the structure in Figure XXV, where MANY is a silent "gradable determiner" (cf. Solt, 2014; Kennedy, 2015).

(5:10) More than five dots are blue.

He defines MANY as shown in 5:11. That is, as a boolean function that takes a degree, d, and two properties over pluralities, P and Q, as arguments and returns 1 iff the number of atomic parts that have both properties is at least d.

(5:11) Hackl's denotation for MANY:
$\textsc{many} := \lambda d.\lambda P.\lambda Q.\exists x\big(P\ x \wedge Q\ x \wedge \#(x) \geq d\big)$

Crucially, as was the case with ordinary comparatives, *fewer* can be decomposed into *-er few*, also in comparative quantifiers. This was worked out in detail by Solt (2009, 2014), for example (cf. also Penka, 2015; Beck, 2012a). A slightly different possibility that is based on Rullmann (1995) and I. Heim (2006) is described in appendix A.2. The integrated processing model that is presented in the next section implements yet another possibility.

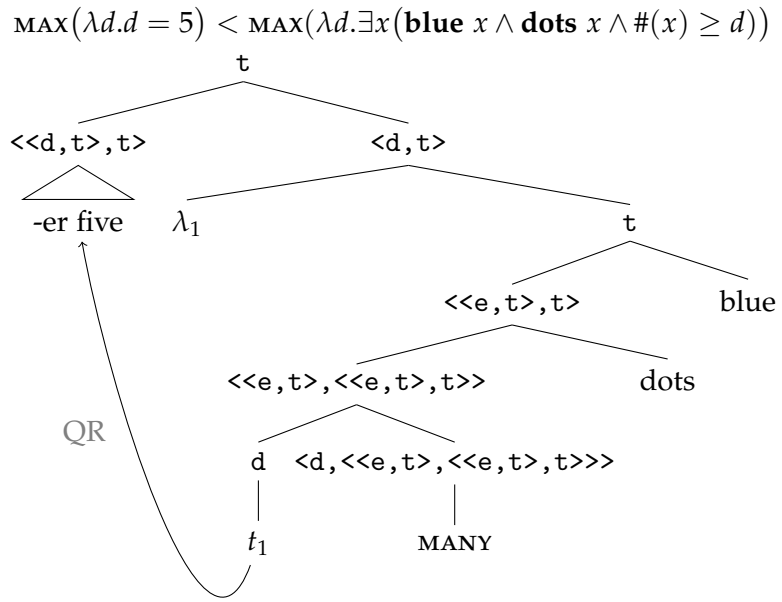


Figure XXV. Logical form proposed by Hackl for the sentence 5:10. For simplicity, *than* is omitted in this tree as it is assumed to be semantically vacuous. Alternatively, it could be interpreted as an identity function of the appropriate type. Moreover, while the auxiliary *are* surely has semantic content (e.g. tense), we can also treat it as vacuous for the present purpose.

We conclude that linguistic analysis gives reasons to believe that there is a semantic operator, FEW, in *fewer than n*, which is absent in *more than n*. Similar to covert negation, this operator may correspond to an extra processing step which could, in turn, explain that the DE modified numerals take longer to evaluate than the UE ones. However, processing models of quantifier verification are not as readily amendable with a processing step corresponding to FEW as they are with covert negation. We noted above that the latter can be added to any processing model describing truth evaluation. In contrast, application of the antonym operator FEW is more restricted. For example, if we assume, with Rullmann (1995), that the semantic contribution of few is scale-reversal, it is not obvious how to amend the semantic automata (section 5.2.1) with an extra processing step corresponding to FEW because they involve no scales. In conclusion, while scale-reversal induced by *few* may be a semantically plausible and intuitively appealing explanation of enhanced processing difficulty, we

still have to formulate a processing model into which a corresponding processing step can be integrated in order to make this explanation explicit. To this end, a sequential sampling model like the one laid out above seems appropriate as one of its basic assumptions was that numerical information is represented on an analog scale.

5.3 AN INTEGRATED PROCESSING MODEL

The semantic theory just outlined can be combined with a sequential sampling model akin to the one discussed in section 5.2.2. Thereby, we derive semantic representations and, importantly, also verification processes compositionally. In the present section, such an *integrated processing model* (IPM) is developed step by step. First, it is shown how the semantic theory outlined in the previous section allows us to derive symbolic meaning representations for simple sentences containing comparative modified numerals. Next, we walk through the symbolic representation of the UE case, involving *more than n*, and enrich it with semantic content. The resulting representations can be thought of as snapshots of the verification process. After that, it is shown how to combine these into a genuine process. Finally, we add the operator FEW in the DE case, containing *fewer than n*. The contribution of FEW is an extra processing step corresponding to scale-reversal. Theoretical and empirical motivation is provided throughout.

Before the IPM is developed, a general comment is in order. Notice that a large number of influential proposals were made within the last fifty years or so that aimed at a unified analysis of comparatives (among many others: Bresnan, 1973; Cresswell, 1976; Klein, 1980; von Stechow, 1984; Kennedy, 1997; I. Heim, 2000; Lechner, 2001; Schwarzchild & Wilkinson, 2002; Neeleman, Van de Koot, & Doetjes, 2004; Bhatt & Pancheva, 2004; Grosu & Horvath, 2006; Beck, 2011; Wellwood, 2015). Some basic assumptions about syntax and semantics differ greatly between the different proposals. The few detailed derivations discussed thus far already hint at this diversity (e.g. section 2.6). It is beyond the scope of the present work to evaluate or compare the different proposals. Rather, the aim is to extract what appear to be the minimally required building blocks for an adequate semantics of comparative modified numerals and at the same time to presuppose as little as possible of any particular syntactic or semantic theory.

Despite the diversity of approaches, the basic semantic building blocks are strikingly uniform. In particular, the assumed lexemes and morphemes are essentially the same across many theoretical proposals and their respective semantic contributions to sentence meaning are also virtually identical. For comparative modified numerals, the following lexical items seem to be required at the minimum, at least in a degree-based approach:

R1: a numeral, which denotes a degree,

R2: some analogue of MANY that introduces a second degree,

R3: the comparative morpheme *-er*, which expresses a comparison between degrees,

R4: an analogue of *few* for the DE case that has the effect of scale-reversal;

In addition to these four requirements, there is arguably a fifth one. This has to do with the fact that comparative sentences may optionally contain further degree expressions. For example, the measure phrase *at least six inches* combines with *taller than Simon* in the following *differential comparative*.

(5:12) Elin is at least six inches taller than Simon.

Similarly, comparative modified numerals can also contain further degree expressions:

(5:13) a. Jogi invited a few more than 50 guests.

b. Many more than 100 students attended the lecture.
(from Solt, 2014)

Differential comparatives are clearly different from the other examples of comparatives, discussed so far. There are several approaches how to accommodate both types of sentences. The standard solution (within the relational approach, see section 2.6.2.1) is to provide the comparative morpheme with an extra degree-argument position for phrases like *at least six inches*. In case no such degree expression is present, existential closure applies to bind this argument position (e.g. von Stechow, 1984). Another, closely related solution is to assume that the comparative morpheme is lexically ambiguous: One

version of *-er* has an additional degree-argument position compared to the other (for discussion see Beck, 2011). Kennedy and McNally (2005a, 2005b) proposed yet a third solution, which is also the one adopted here. They recognized that comparatives behave parallel to what is observed with bare gradable adjectives. These may also optionally combine with overt degree expressions like measure phrases, for example. In case they do not – referred to as the *positive form* – it is standardly assumed that they combine with a silent operator called POS (Cresswell, 1976; von Stechow, 1984; Kennedy, 2007; see also section 2.6 for technical details). This makes a uniform lexical semantics for gradable adjectives possible. Kennedy and McNally propose that the same mechanism is operative in comparatives. More specifically, they propose that combining a gradable adjective, the comparative morpheme *-er* and a *than*-phrase yields an expression (e.g. *taller than Simon*) that again has the semantic type of a gradable adjective. In their own terminology: "(unmodified) comparative constructions are semantically derived minimum-standard absolute adjectives" (Kennedy & McNally, 2005a, p. 374; cf. also Svenonius & Kennedy, 2006). This brings us to our fifth requirement:

R5: the POS operator if no overt degree expression is present.

As compared to the first four, there seems to be more room for debate regarding this one. While the problem is an obvious one, the latter solution is not the standard one. With regard to *few* and *little*, it is, however, an attractive one because assigning identical semantic types to both, bare gradable adjectives and bare comparatives allows for these operators to compose with both types of expressions just in the same way. In other approaches this needs complicated machinery.

5.3.1 *Deriving symbolic meaning representations*

Based on the introduced assumptions about the semantic building blocks of comparative modified numerals, we can derive the symbolic meaning representations in 5:14-b and 5:15-b for the sentences in 5:14-a and 5:15-a, respectively.

(5:14) a. More than five dots are blue.
b. POS (ER (MANY **dots blue**) **five**)

(5:15) a. Fewer than five dots are blue.
b. POS (FEW (ER (MANY **dots blue**) **five**))

This is demonstrated here using the *combinatory categorial grammar* (CCG) formalism (Ades & Steedman, 1982; Szabolcsi, 1987, see also section 2.7). The main reason that this formalism was chosen here is that it allows us to derive the desired representations from the assumed building blocks without the need for additional, construction-specific assumptions. In the next section the derived representations allow us to formulate a transparent interface to truth evaluation processes – in line with the ITT.

A derivation of 5:14-b is shown in Figure XXVI. The top row corresponds to lexical retrieval. The dotted line stands for lexical decomposition. In line with our assumptions, the word *more* is decomposed into POS, *-er* and MANY. How lexical decomposition works in detail is left open here. Some morphological mechanism is stipulated that mediates between the surface form *more* and its possible semantic decompositions. After decomposition, the involved lexemes can interact with other parts of the sentence. However, derivation rules only apply to adjacent items. The syntactic categories are in line with our five requirements. There are different possibilities to derive 5:14-b from 5:14-a using these categories. The one shown here first combines POS, *-er* and MANY yielding the syntactic category of the word *more*. The derivation proceeds more or less incrementally. The representation in 5:15-b is derived analogously to 5:14-b. The only difference is that *-er* combines with *few* before it combines with MANY. This sub-derivation is shown in Figure XXVII. The syntactic category of *few* is $\text{DEG}_{adj}|\text{DEG}_{adj}$. The vertical slash is used to indicate that *few* 'is looking' for a degree expression to its left or to its right.

Two more comments are in order. Firstly, the derivation in XXVI makes rudimentary use of *features* to rule out ungrammatical derivations. For example, the fact that the comparative morpheme selects for a *than*-phrase is encoded in the category DEG_{than}. Secondly, the operator POS has a subscript. The idea is that the implicit standard of comparison depends on the function in this subscript. This is discussed briefly in section 2.6. There, it is also explained that, in addition, the standard depends on a contextually determined comparison class. These points are important for gradable adjectives in the positive form but less relevant for comparatives. Like in 5:14-b and 5:15-b, the subscript of POS is therefore often omitted in what follows.

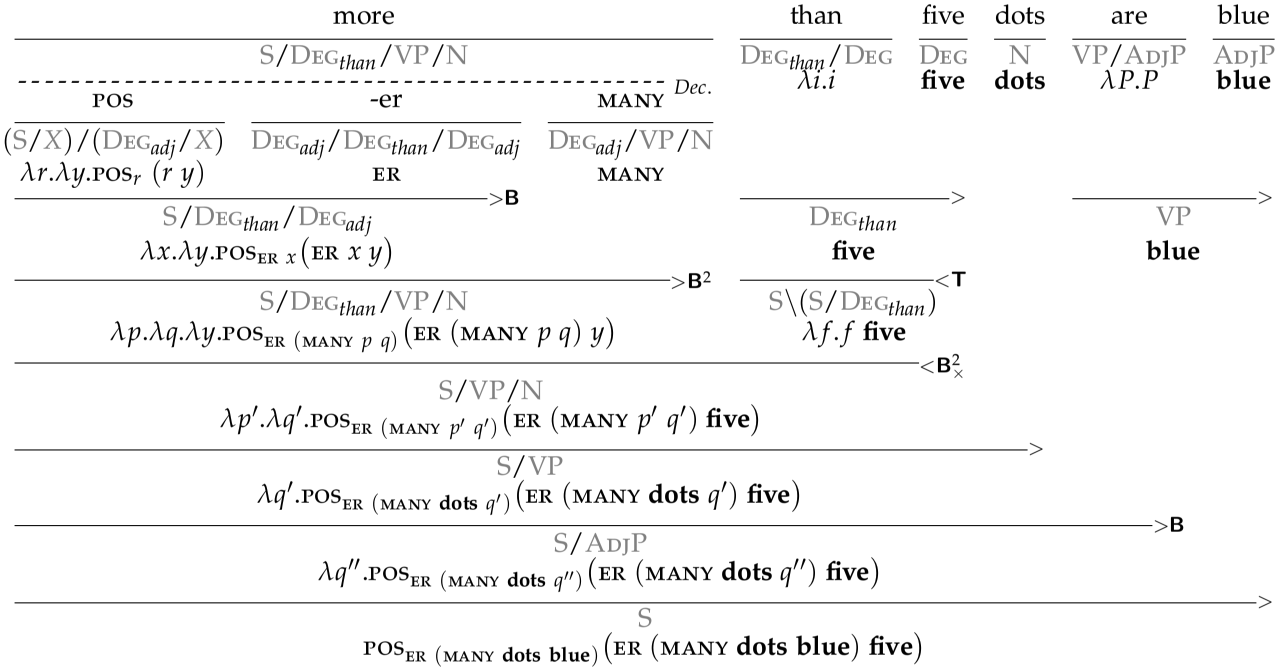


Figure XXVI. CCG derivation of the example in 5:14

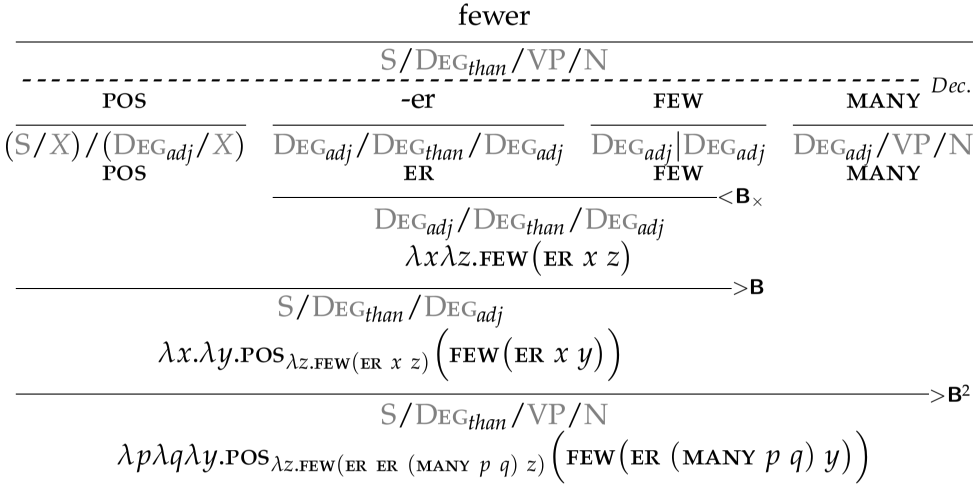


Figure XXVII. Subderivation needed for 5:15.

5.3.2 *Developing the truth evaluation process*

Next, we map the symbolic meaning representations to a truth evaluation process. We start with the UE case: POS (ER (MANY **dots blue**) **five**)

5.3.2.1 *Representation of* MANY **dots blue**

In a situation in which, for example, m blue dots are visually presented, MANY **dots blue** encodes a representation of m. That is, MANY is interpreted as a function that maps two properties to a representation of how many objects have both properties (cf. Hackl, 2000, 2002). This representation assigns real numbers to points on a scale. In particular, m is represented by a function f_m (cf. Dehaene, 2007) with:

$$f_m(x) := \begin{cases} \alpha_m \phi\left(\frac{log(x)-log(m)}{\sigma}\right), & \text{if } x > 0 \\ 0, & \text{otherwise,} \end{cases} \tag{5:16}$$

where x corresponds to the position on the scale, σ and α_m are constants, and ϕ is the Gaussian function with:

$$\phi(x) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right).$$

This kind of representation has played an important role in theories of numerical cognition. It was implemented in Dehaene and Changeux's (1993) influential connectionist model of basic numerical abilities. Their model was based on 'number detector neurons' which had activation functions as in equation 5:16 (see also Verguts & Fias, 2004, for an extension of the model). In particular, when the numerosity m was presented to the network, the activation of detectors for k was approximately $f_k(m)$. As a consequence, the numerosity m was represented according to f_m in the population of number detectors: The activation over a sequence of detectors, $(d_i)_{i\in\{1,\ldots N\}}$, where each d_i has activation function f_{θ_i}, is captured by this function. If m is encoded, then $d_i = f_{\theta_i}(m) = f_m(\theta_i)$.

Empirical support for the existence of such detectors and the corresponding representations was provided by Nieder and Miller (2004) and Nieder and Merten (2007) from single cell recordings in the posterior parietal and prefrontal cortex of rhesus macaques. Number selective neurons were identified which had average firing rates obey-

ing the main characteristics of equation 5:16: They followed Gaussian curves with constant width when plotted on a logarithmic scale. Moreover, the fMRI study by Piazza et al. (2004) provides evidence for the hypothesis that the same encoding of numerosities is employed by humans.

The function f_m may be thought of as a crude approximation of a *population code* for m (i.e. a pattern of neural activity that encodes some stimulus or quantity; for details see, e.g., Dayan & Abbott, 2001 or Mallot, 2013). The right panel of Figure XXVIII shows the representations for $m \in \{3, \ldots, 8\}$. The maxima of the curves are equally spaced, but the curves are asymmetric and their widths increase with m. Realistically, population codes are noisy and thus, in one particular time window, may look more like is sketched in the right panel of the figure. For simplicity, I ignore noise for now but come back to this important point in subsection 5.3.2.5. Furthermore, I use functions from $\mathbb{R}$ to $\mathbb{R}$, which Dayan and Abbott (2001) refer to as *continuously labeled* population codes. However, all of what follows can, in principle, also be done with discrete sequences.

There is an indirect connection between these representations and the log-Gaussian model of number representations discussed above (subsection 4.2.2.1, especially Hypothesis 4:16). Dehaene (2007) suggested to link the neuronal and the psychophysical level via a *bridging law*: He assumes a population of detector neurons that have activation functions as in 5:16 but are also subject to noise. The psychophysical, log-Gaussian model is then derived via a statistical estimator that infers the location on the mental number line from the population code. Bridging laws like this are sensitive to a number of factors including the assumed activation functions, the kind of noise, the used estimator and assumptions about how the detectors are distributed over the number line, so to speak (Snippe, 1996; Dayan & Abbott, 2001). Here, I do not apply a bridging law of this kind but work directly with f_m as defined in 5:16.

5.3.2.2 *Representation of the numeral*

Next, we consider the representation of the numeral. We assume that it is represented by a function g_n with:

$$g_n(x) := \frac{1}{\sigma'} \phi \left(\frac{x - n}{\sigma'} \right), \tag{5:17}$$

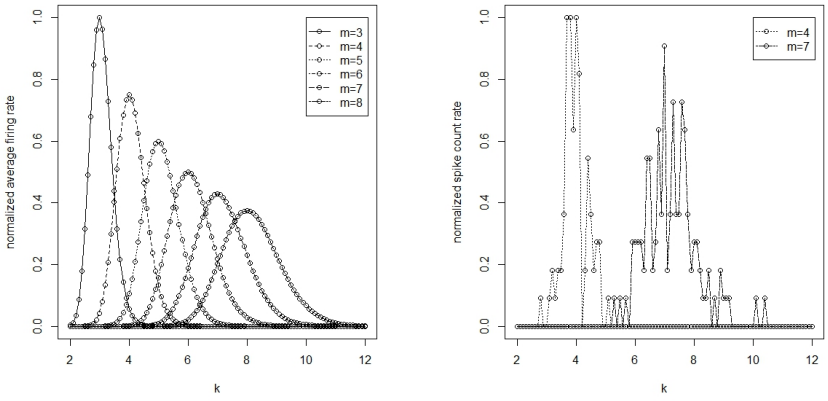


Figure XXVIII. Sketch of population codes for numerosity.

where ϕ is the Gaussian function introduced above (e.g., **five** $:= g_5$). The parameter σ' is assumed to be very small. In contrast to the numerosity representations computed by MANY, the numerals are thus represented by very narrow bell-shaped curves. This provides a precise and almost exact encoding. Another difference is that the representations of symbolic numerals are symmetrical whereas the numerosity representations are not.

These assumptions are in accordance with Dehaene's (1997) hypothesis that symbolic number representations reuse the representation format of approximate numerosities but have greater precision (see also Dehaene, 2007; Dehaene & Cohen, 2007; Nieder & Dehaene, 2009).

Hypothesis 5:18. *The meaning representations of number symbols and numerals reuse representations of approximate numerosity. In addition, the processes that operate on theses representations are also reused.*

This hypothesis is supported by a number of empirical findings (for discussion see Dehaene, 2007, p. 551–561). Additional evidence comes from the simulation study of Verguts and Fias (2004), who first trained an artificial neural network to represent numerosities and then to represent symbolic numerals (see section 5.6.2, below for discussion).

5.3.2.3 *Contribution of* ER

Since the expressions discussed so far received non-standard semantics, it does presumably not come as a surprise that the compara-

tive morpheme *-er* also receives an unusual one: The operator ER computes a representation of the difference between the quantities encoded by **five** and MANY **dots blue**. This is accomplished by the following computation, called cross-correlation (symbolically, $\star$):

$$(\text{ER } f\ g)(x) := (g \star f)(x) = \int_{-\infty}^{+\infty} g(\tau) f(x+\tau) d\tau. \tag{5:19}$$

If ER is applied to g_n and f_m, for some n and m, we can think of this computation as a translation of f_m in the negative direction. This is because we assumed σ' to be very small and the following fact holds (see e.g. Amann & Escher, 2001; Weber & Arfken, 2003):

$$\lim_{\sigma' \to 0} (g_n \star f_m)(x) = f_m(x+n). \tag{5:20}$$

As σ' approaches 0, $g_n \star f_m$ converges to a shifted version of f_m which is translated an amount of n in the negative direction. In that sense, $g_n \star f_m$ represents the difference between (an approximately represented) m and (an almost exactly represented) n. I use cross-correlation here instead of directly defining ER as translation because ER should also be applicable to two approximate representations (e.g. f_n and f_m). An obvious example application would be the case where a comparative sentence like *more dots are red than green* is evaluated under time pressure in situations with lots of dots (as was studied in an experiment by Deschamps, Agmon, Loewenstein, & Grodzinsky, 2015, see section 6.2.2 for discussion).

Interestingly, Pouget, Deneve, and Duhamel (2002) proposed a neural implementation of this kind of computation as a model of how neural representations are transformed during multisensory integration (see also Dayan & Abbott, 2001; Pouget & Sejnowski, 1997). An example is the transformation from eye-centred coordinates (visual location) to head-centred coordinates (auditory location) given the viewing angle of the eye. Such transformations shift the frame of reference. The model of Pouget et al. (2002) accounts for a range of empirical data about neural activity involved in multisenory integration. In consequence, the kind of computation that is used here to model the contribution of the comparative morpheme *-er* is plausible from a neurobiological perspective. Moreover, Lipinski, Schneegans, Sandamirskaya, Spencer, and Schöner (2012) use similar computations in their model of spatial prepositions like *under* or *over*. In their model,

a representation of the relative position of a "located object" relative to a "reference object" is computed.

Putting the components together, we have:

$$\textsc{er}\ (\textsc{many}\ \textbf{dots blue})\ \textbf{five} = g_5 \star f_m,$$

which is a representation of $m - 5$. Of course, this does not yet define a truth value. However, truth-value judgments can be based on this representation. This is where POS comes in.

5.3.2.4 *The operator* POS

The operator POS does two things. Firstly, it introduces a threshold for the difference $m - n$ to surpass (cf. Svenonius & Kennedy, 2006 for linguistic motivation). I assume this threshold to be set at 0.5 for the case of comparative modified numerals. Thus, the "yes, true" response is favored in a verification task if the difference between m, the number of target objects and n, the numeral mentioned in the sentence surpasses 0.5. This thresholds seems reasonable for the case were numbers and cardinalities are compared, especially since they are assumed to be represented in an noisy and analog format.

Secondly, POS weighs evidence for the "yes, true" response against evidence for the "no, false" response. In particular, the following quantity is computed:

$$\textsc{pos}(h) := \int_{0.5}^{\infty} h(x)dx - \int_{-\infty}^{0.5} h(x)dx. \tag{5:21}$$

If the area under the graph of h that is to the right of 0.5 is larger than the area to the left of 0.5, $\textsc{pos}(h)$ is positive and a "yes, true" response is favored. In contrast, a "no, false" response is favored if $\textsc{pos}(h)$ is negative. For large absolute values, one response alternative clearly outweighs the other one. Based on the characteristics of $g_n \star f_m$, specifically its asymmetry, we may expect size and distance effects (see subsection 5.2.2). We did, however, not define a decision process, yet. This is what we turn to next.

5.3.2.5 *The decision process*

Up to now, we have considered static representations. Now, it is sketched how these can be combined into a dynamic verification process. Imagine a situation in which, first, a sentence of the form *more than n dots are blue* is read and then a picture with m blue dots is in-

spected. The task is to decide as fast and accurate as possible whether the sentence is true regarding the picture or not.

While the sentence already introduces the semantic representation of the numeral, the number of target objects cannot be processed before the picture is first inspected. This point in time is denoted by t_0. Making a few simplifying assumptions, we can describe the truth evaluation process from t_0 onwards based on the introduced representations and computations. In particular, we assume that after some time delay, t_{num}, the representation of the number of target objects, m, namely MANY **dots blue** = f_m, has 'built up'. Earlier, it was mentioned that, realistically, this representations and also that of the numeral are corrupted by noise. For simplicity, we still ignore noise at this point and postpone the addition of noise to the last step of the process. Moreover, we assume that representations remain static once they have built up.

The two representations **five** $= g_5$ and MANY **dots blue** $= f_m$ are fed to ER. Again, some time, t_{comp}, passes for the representation $g_5 \star f_m$ to emerge. Afterwards, this representation is passed to POS. To describe the decision process, we make two more assumptions. Firstly, that the value computed by POS is accumulated over time in order to form a 'decision variable' X_t. Secondly, we add noise to X_t. A natural way to incorporate noise is to add a *Wiener process* W (see section 2.5.2.2). For $t_{start} := t_0 + t_{num} + t_{comp}$ and c some constant that determines the amount of noise, this yields a *drift diffusion process* (also defined in section 2.5.2.2) which is described by the following *stochastic differential equation*:

$$dX = (\text{POS}(\text{ER}\ f_m\ g_n))\ dt + c\ dW, \quad X_{t_{start}} = 0. \tag{5:22}$$

This is an instance of Ratcliff's (1978) DDM. A thorough analysis of this process and of connectionist models implementing (or approximating) it was conducted by Bogacz et al. (2006). In section 2.5.2, it is briefly discussed how the time continuous processes as in 5:22 relate to time discrete sequential sampling models.

SOME THEORETICAL JUSTIFICATION Suppose we observe a sequence, $(y_i)_{i\in\mathbb{N}}$, of realizations of i.i.d. random variables $(Y_i)_{i\in\mathbb{N}}$. Moreover, assume that the latter constitute the mental representation of some stimulus in an experiment. Our task is to decide, based on the sequence of realizations, from which of two sets the stimulus was

drawn. We can in principle think of this task as the following hypothesis testing problem (recall technical details from section 2.5.2):

$$H_0 : Y_k \sim P_0, \qquad k = 1, 2, \ldots$$
$$H_1 : Y_k \sim P_1, \qquad k = 1, 2, \ldots.$$

Then, as we have seen, given an appropriate stopping and decision rule, an optimal decision process can be described as a random walk that accumulates the LLR. Moreover, this process converges to a drift diffusion process as in 5:22 if more and more observations are sampled per time interval. However, in order to carry out the optimal decision procedure, the corresponding densities p_1 and p_2 have to be known and, as discussed by Platt et al. (2008), this is a non-trivial obstacle.

Interestingly, Gold and Shadlen (2001, 2002) have shown that in 2AFC tasks – with two alternatives A_1 and A_2 – neural computations can, under several plausible circumstances, approximate the optimal decision processes even without knowledge of p_1 or p_2. Suppose that the $y_1, y_2, \ldots$ represent firing rates of some neural computing unit that responds strongly if response alternative A_1 is correct but responds only weakly if the other alternative A_2 is the correct one. Furthermore, suppose that $z_1, z_2 \ldots$ represent the activity of an "opposing" computing unit with the opposite response profile, A_1: weak response, A_2: strong response. Then, in several plausible scenarios (in particular different types of conditional distributions of Y_i and Z_i given A_1 or A_2 is correct), the difference $y_i - z_i$ is proportional to the change in the LLR. Therefore, accumulating the difference in firing rate between two such units can approximate the optimal decision process. Gold and Shadlen argue that this kind of heuristic is exploited by neurobiological mechanisms that underly decision making.

Concerning the truth evaluation of modified numerals, these considerations can be taken as theoretical justification of the decision process proposed here if we interpret the value computed by POS along the lines just sketched, namely as the difference between two such opposing computing units (or pools of computing units). Moreover, notice that, if we interpret the value computed by POS along these lines, only weak assumption are needed in order for the process in 5:22 to be an adequate description. In particular, if we assume that

during the decision process samples from a distribution with mean POS(ER f_m g_n) and variance c^2 are accumulated, then the process in 5:22 is an adequate description in the limit (cf. section 2.5.2 and references mentioned there). This is the justification that we ignored noise above.

5.3.2.6 *The operator* FEW

The operator FEW is defined as follows (cf. Rullmann, 1995 and I. Heim, 2006) :

$$(\textsc{few}\,(h))\,(x) := h\,(-x)$$

In the DE case, the drift (or rate of accumulation) in equation 5:22 would thus be POS $(\textsc{few}\,(\textbf{er}\ f_m\ g_n))$. It is assumed that FEW corresponds to an additional processing step. Some additional time, t_{few}, is needed for this step to complete. In consequence the decision process starts with a delay, such that, in the DE case, $t_{start} := t_0 + t_{num} + t_{comp} + t_{few}$.

5.3.2.7 *Upshot*

On the basis of symbolic meaning representations, we have characterized truth evaluation processes. The processes for the UE and the DE case are depicted in the left and right part of Figure XXIX, respectively. Because an extra processing step is incorporated in the DE as compared to the UE case, the model accounts immediately for the observed prolonged RTs of the DE comparative modified numerals. In addition, it is compatible with findings from numerical cognition like the often observed size and distance effects.

In developing the model, a number of simplifying assumptions were made. For example, a serial process was assumed, in which one processing step has to complete before the subsequent one starts (Sternberg, 1969). Of course, models with continuous, partial outputs of processing stages are conceivable as well (cf. McClelland, 1979). Furthermore, some aspects or parameters are left entirely unspecified. For example, it is left completely unspecified how the representation of the number of target objects is arrived at. Several alternatives are conceivable in this regard. The two extremes are 'pure approximation' and 'exact counting'. Moreover, there are a number of model parameters that are left unspecified here but would be crucial to derive quantitative predictions. Because of all this underspecification, quan-

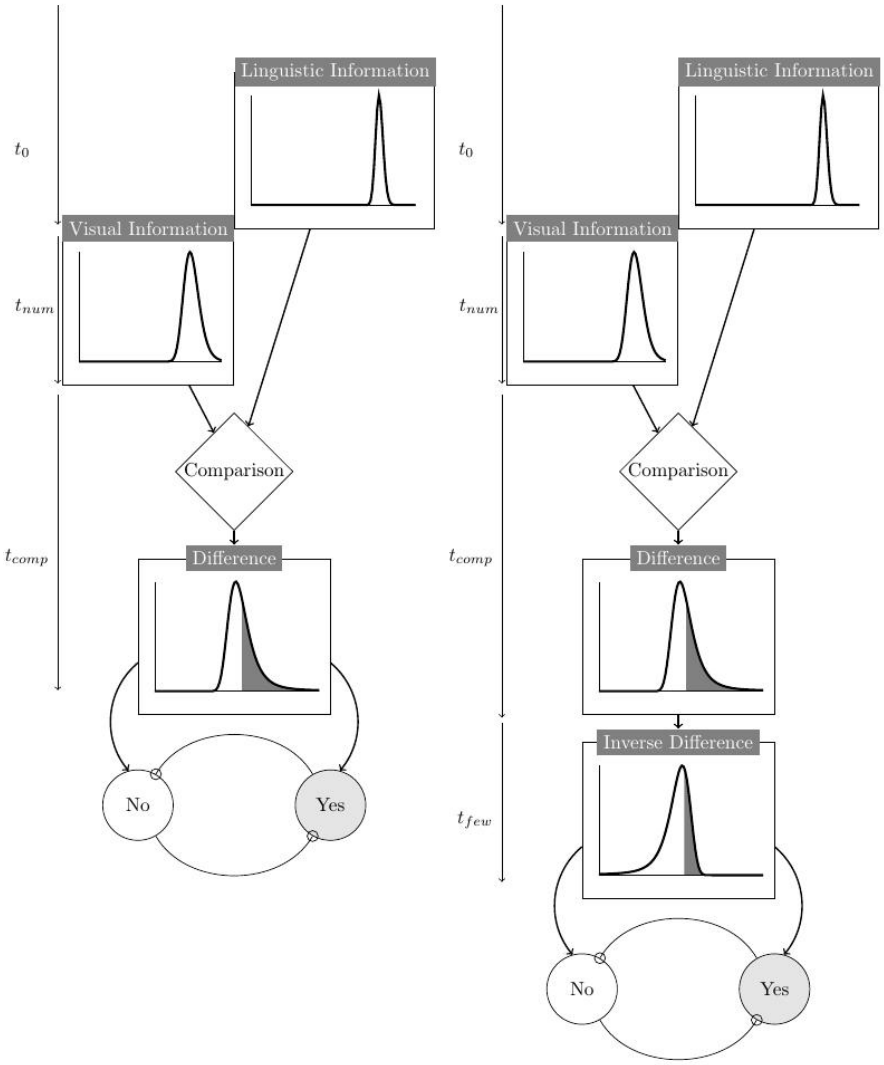


Figure XXIX. Graphical representation of truth evaluation processes.

titative predictions are not possible or intended here. Nevertheless, the model allows for some qualitative predictions. In the following these will be discussed and tested.

5.4 EXPERIMENT 3A: ORDINARY SENTENCE-PICTURE VERIFICATION

Two types of potential accounts for prolonged judgment times of *fewer than n* as compared to *more than n* were identified in the previous sections. Under one type of account, the increased difficulty is attributed to an additional semantic operation in the former as compared to the latter case. Several alternatives of how this can be

spelled out were sketched. The other type of account assumes no additional operators and also no inherent difficulty of the DE case but attributes the observed effects to the processing of numerical information involved in the verification task. It was shown in section 5.2.2 that, under defensible assumptions, such effects would indeed be expected even if the direction of entailment is ignored completely.

The purpose of the present experiment was twofold. Firstly, it was intended to replicate the prolonged RTs for *fewer than* as compared to *more than n* in a slightly different experimental design than what was employed in previous studies. Secondly, it was intended to decide between the two types of accounts just mentioned. More specifically, the experiment was designed in such a way that theories of numerical cognition would – to the extent that they let us expect any difference at all – predict *more than n* to take longer to evaluate than *fewer than n*. As a consequence, longer RTs of the latter provide strong reason to focus on the inherent semantic complexity of the DE conditions, in particular, on theories that posit additional semantic operations in these types of linguistic constructions.

Participants had to evaluate sentences containing *more than n* or *fewer than n* against pictures showing *n* target objects. The number of depicted objects and the correct judgment (namely "no, false" in both cases) did not differ between the UE and the DE quantifiers and, at the same time, identical picture materials were used with both types of quantifiers. The only difference between the stimulus material in the DE and UE conditions was the modifier: *more* vs. *fewer*.

5.4.1 *Method*

5.4.1.1 *Participants*

Fourty-eight participants were recruited at the University of Tübingen (30 female). Their mean age was 25.6 (ranging from 20 to 33 years). All participants had normal or corrected to normal vision. Participants were naïve to the purpose of the study. They received a small financial compensation.

5.4.1.2 *Materials*

Sentence materials consisted of German sentences like in 5:23. The factor *modifier* (two levels: *more* and *fewer*) was crossed with the factor *numeral* (four levels: *four*, *six*, *eight*, *ten*) yielding the eight conditions

exemplified in 5:23-a-h. Sentences in all eight conditions contained a color adjective (*blau* ('blue'), *rot* ('red'), *grün* ('green') or *orange* ('orange')) and a noun describing a shape (*Punkte* ('dots'), *Quadrate* ('squares'), *Dreiecke* ('triangles') or *Kreuze* ('crosses')). There was a locative PP and an auxiliary between the noun and the adjective.[6]

(5:23) a. Mehr als vier Punkte auf dem Bild sind blau.
More than four dots on the picture are blue.

b. Weniger als vier Punkte auf dem Bild sind blau.
Fewer than four dots on the picture are blue.

c. Mehr als sechs Punkte auf dem Bild sind blau.
More than six dots on the picture are blue.

d. Weniger als sechs Punkte auf dem Bild sind blau.
Fewer than six dots on the picture are blue.

e. Mehr als acht Punkte auf dem Bild sind blau.
More than eight dots on the picture are blue.

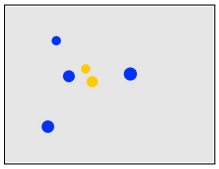
f. Weniger als acht Punkte auf dem Bild sind blau.
Fewer than eight dots on the picture are blue.

g. Mehr als zehn Punkte auf dem Bild sind blau.
More than ten dots on the picture are blue.

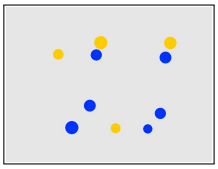
h. Weniger als zehn Punkte auf dem Bild sind blau.
Fewer than ten dots on the picture are blue.

Sentences were paired with pictures that depicted randomly distributed shapes of two colors. Example pictures are provided in Figure XXX. Each picture showed objects of the shape mentioned in the corresponding sentence. One of the two colors in each picture matched the color adjective in the corresponding sentence. If this color was blue, the remaining objects were orange; if it was red, the remaining objects were green; and *vice versa*. Depending on the numeral in the sentence, pictures showed four, six, eight or ten target objects (e.g. blue dots). The number of target objects always matched the numeral mentioned in the sentence and sentences never described their corresponding picture truthfully. One third of the pictures showed a number of non-target objects that was smaller than the number of target objects. Another third showed exactly as many non-target as target objects. The remaining pictures had a number of non-target objects that was larger than the number of target objects. Within the pictures that showed different numbers of target

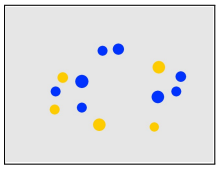
[6] Eye-movements during reading were recorded in order to test for increased processing difficulty in sentences with DE as compared to UE quantifiers. These are not reported here but in chapter 6. The locative PP served as a buffer region.



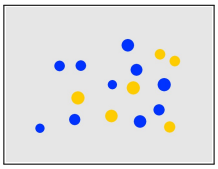
(a) 4 targets, 2 non-targets



(b) 6 targets, 4 non-targets



(c) 8 targets, 5 non-targets



(d) 10 targets, 6 non-targets

Figure XXX. Example pictures from Experiment 3

and non-target objects, the ratio of the larger and the smaller number was approximately constant.

In total, 320 such pictures were generated – 80 for each of the four numerals. Each picture was paired once with a sentence containing the modifier *more* and once with a sentence containing *less*. Thus, there were 640 sentence-picture pairs in sum, which comprised 80 experimental items. Each item consisted of eight sentences combined with four pictures. Within each item, sentences with the same numeral also were paired with the same picture. In consequence, the only difference between UE and DE conditions was the modifier, i.e. the first word of the sentences.

The sentence-picture pairs were distributed over eight lists using a Latin square design. This ensured that each participant saw each item – and thus each picture – only once. Each list contained 80 experimental trials. In addition, 343 filler trials were included. These were sentence-picture pairs similar to the experimental items but possibly containing different quantifiers or negation. About half of the trials in each experimental list required a "yes, true"-judgment and the other half required a "no, false"-judgment.

5.4.1.3 *Procedure*

The experiment was conducted using an Eyelink 1000 eye-tracker and Eyelink's Experiment Builder software. However, eye movements are not reported here (they are discussed in chapter 7). Participants were tested individually in a silent room. They sat about 70 cm away from a 27 in computer screen and positioned their head on a chin and fore-

head rest. While participants read the stimulus sentences, the position of their dominant eye was sampled with a frequency of 1000 Hz. At the beginning of the experiment, they received written instructions on the screen. After they had read the instructions, the eye tracker was calibrated using a nine point grid. It was ensured that fixations were tracked within 0.5° of visual angle across the entire area of the screen. After the calibration, participants completed ten practice trials, in which feedback was provided to them. The rest of the experiment consisted of four blocks. Participants were told that they could rest between blocks. An experimental session took around one hour.

At the beginning of each trial, there was a calibration check. Participants had to fixate a cross in the upper left portion of the screen. A recalibration was launched if no fixation was registered within a radius of approximately 1.5° of visual angle around the fixation cross for a time period of 10 s. After a fixation was registered, the fixation cross disappeared and a sentence was presented. The first word of the sentence was centered where the fixation point had been presented. Three characters spanned approximately 1° of visual angle. Participants read the sentence and afterwards fixated a cross in the lower right portion of the screen. They were instructed to read fast but at a pace that allowed them to understand the sentences.

After reading the sentence, a picture was presented to the participants in central position on the screen, spanning approximately 30° of visual angle. They provided a truth-value judgment by pressing one of two buttons on a joypad. They were instructed to respond fast and accurately. There was a time limit of 13 s. The response-key assignment was counterbalanced between participants. There were two version of each experimental list with opposite response-key assignments. RTs, i.e. the time between the onset of the picture presentation and the judgment, were recorded.

5.4.1.4 *Predictions*

The automata model (section 5.2.1) amended with negation as an additional processing step in the DE conditions (section 5.2.3.1) predicts that these conditions take a constant amount of time longer to evaluate than the UE conditions. The extra time is needed to flip the truth value. Apart from this effect, the RTs should increase with the number of objects that have to be counted. In both the UE and the DE conditions, the number of objects to count is determined by the numeral in the sentence.

Under the sequential sampling model based on log-Gaussian representations (section 5.2.2), the RTs consist of two components: the decision and the non-decision time. The mean decision time is predicted to be proportional to $log((n+0.5)/n)$ and $log((n-0.5)/n)$ in the UE and DE case, respectively. Since $log(n+0.5/n) < log(n-0.5/n)$ and non-decision time is not expected to differ, the mean RT is predicted to be shorter in the DE than in the UE conditions (if the subtle difference is expected to produces any difference at all). Similarly, the proportions of errors should also be higher in the UE case.

Just as the automata model, the sequential sampling model can also be amended with an additional processing step corresponding to negation. This predicts an increase in the non-decision times in the DE conditions. In this scenario, predictions about mean RT are ambiguous. The extra processing step prolongs the RTs of the DE conditions since they have longer non-decision times. In contrast, the decision times are predicted to be faster in the DE than the UE conditions. Despite the ambiguity, longer RTs of the DE as compared to the UE conditions would constitute evidence for an additional processing step. The proportion of errors is not expected to be affected by the additional step.

The IPM (section 5.3) combines a sequential sampling model of the decision process with a compositional semantics for comparative modified numerals. Its qualitative predictions are essentially identical to the ones just mentioned. Based on the decision component of the model, the UE conditions are expected to be more difficult than the DE ones. This prediction is derived from asymmetric representations of the number of target objects, which leads to a smaller drift rate (or signal-to-noise ratio) in the UE than the DE conditions. Proportions of errors are therefore predicted to be higher in the UE than the DE conditions. Predictions about RTs are again not unequivocal because an additional processing step is assumed in the DE conditions, which may outweigh the faster decision process in these conditions.

5.4.2 *Results*

Before statistical analysis, contaminated trials were identified and removed. After visual inspection of the fixations during reading, all trials with total reading times below 560ms (103 trials) or above 10s (156 trials) were removed. In total, this affected 1.29% of the trials.

Table XXXI

Descriptive statistics of Experiment 3

numeral	mean proportions of correct judgments (sd)		mean RTs (std. error) in ms	
	fewer	more	fewer	more
four	0.96 (0.07)	0.96 (0.08)	1445 (105.3)	1364 (111.4)
six	0.91 (0.10)	0.92 (0.11)	2124 (134.5)	2097 (165.0)
eight	0.93 (0.11)	0.87 (0.13)	2935 (220.3)	2864 (211.9)
ten	0.88 (0.15)	0.82 (0.16)	3514 (251.2)	3375 (274.7)

Note. All values calculated on the basis of by-subject means.

Moreover, for each condition, trails with RTs that were at least three standard deviations above the mean RT in that condition were removed. Furthermore, trials with RTs below 200*ms* were also removed. The latter two steps affected 1.8% of the remaining trials.

Proportions of correct judgments are summarized in the left part of Table XXXI. They were close to ceiling for the numeral *four* but decreased with the size of the numeral. The decrease was steeper for UE *more than n* than for DE *fewer than n*. A logit mixed effects model analysis was conducted in order to analyze the proportions of correct judgments. First, a saturated model was fit using the `lme4` package (Bates, Mächler, et al., 2015) for the statistical software `R` (R Core Team, 2016). The factors *numeral* (levels: *four*, *six*, *eight* and *ten*) and *modifier* (levels: *more* and *fewer*) as well as their interaction were included as fixed effects. In addition, a random intercept of participants was included as random effect (addition of random slopes led to non-convergence). After that, a model comparison procedure based on the LRT was used to test for significant effects. A significant interaction between *modifier* and *numeral* ($\chi^2(3) = 7.89, p = .048$) was found by comparing the saturated model to a simplified model without the interaction term. The interaction may possibly be due to the fact that the modifier *fewer* led to more errors than *more* in combination with the numeral *six* whereas with the other three numerals UE *more* led to more errors. Next, the procedure suggested by Levy (2014) was used to test for the two main effects in the presence of the interaction. A significant main effect of *modifier* ($\chi^2(1) = 4.625, p = .032$) was found by comparing the saturated model to a simplified one without the factor *modifier*. In addition, the main effect of *numeral* also turned out to be significant ($\chi^2(3) = 70.69, p < .001$) by comparison between the saturated model and a model that did not include this factor.

As shown in the right part of Table XXXI, larger numerals required longer RTs on average. In addition, the DE conditions took longer to evaluate than the UE conditions across all of the four numerals. The RTs were statistically analyzed using linear mixed effects models. As for the proportions of judgments, a saturated model was computed first. The saturated model included the factors *numeral* (levels: *four*, *six*, *eight* and *ten*) and *modifier* (levels: *more* and *less*) and their interaction. In addition, random intercepts and random slopes of *numeral* and *modifier* were included for participants (addition of random intercepts for the interaction led to non-convergence). Again, a model comparison procedure was carried out. There were significant main effects of the *numeral* ($\chi^2(3) = 112.7, p < .001$) and *modifier* ($\chi^2(1) = 3.911, p = .048$). The former reflects the fact that RTs increased with the numeral while the latter was due to longer RTs in the DE as compared to the UE conditions. The interaction was far from significant ($\chi^2(3) = 2.314, p = .510$).

5.4.3 *Discussion*

Truth evaluation of DE *weniger als n* (*fewer than n*) took longer than it took for its UE counterpart *mehr als n* (*more than n*). This replicates previous findings. In addition, the UE conditions led to more errors than the DE ones. For the specific design of the present experiment, where the number of target objects was equal to the numeral in the sentence, the latter effect was predicted by the IPM proposed in section 5.3. Both the IPM and the log-Gaussian sequential sampling model (subsection 5.2.2) amended with covert negation are compatible with the experimental results. In both accounts, the increased proportions of errors in the UE conditions are due to the drift rate (or signal-to-noise ratio) of the decision processes. In addition, the prolonged RTs of the DE conditions are accounted for by an additional processing step. In the IPM, this step consists in scale-reversal induced by the semantic operator FEW whereas in the latter model it corresponds to negation. The only unexpected result is that *fewer than six* led to slightly more errors than *more than six*. I have no explanation for this finding and consider it spurious. Especially since proportions of errors were close to floor in these conditions, a small number of 'fast guesses' may have caused it.

Even in an experimental design where the numerical comparison should be more difficult in the UE conditions, the DE conditions took

longer to evaluate than the UE ones. This supports theories that posit inherent difficulty of *fewer than* as compared to *more than*. Specifically, the experimental results support theories that assume an additional semantic operation in the *fewer than* that does not take part in the evaluation of *more than*. More generally, the results pose non-trivial constraints on possible processing theories of the truth evaluation of comparative modified numerals because such theories should account for both proportions of errors and RTs.

The semantic automata model is at odds with the experimental results. Even if it is amended with an additional operation corresponding to negation, the automata model does not account for the increased proportions of errors in the UE conditions. In the semantic automata model, no errors are predicted at all. A perspective that may be interesting in this regard are the probabilistic automata proposed by Dotlačil et al. (2014) that was already discussed briefly in section 4.1. However, it is not clear to me what this model would have predicted for the present experiment. Similarly, the log-Gaussian sequential sampling model also provides no immediate account of the experimental results on its own, without an additional processing step in the DE conditions.

However, there is a possibility to explain the findings without the assumption of an additional processing step. Assume, for the moment, that evaluation of *fewer than n* poses the same difficulty as evaluation of *more than n*, but participants perceive *fewer than n* to be more demanding. A potential reason why evaluation of *fewer than n* could be perceived as demanding is its more complex compositional structure. Under these assumptions, the results could be explained in terms of a speed-accuracy tradeoff (SAT): Because the DE conditions are perceived to be more difficult, participants could have decided to invest more time in these conditions, which, in turn, would have resulted in higher accuracy. In a sequential sampling model, this could be modeled by response boundaries that are spread further apart in the DE as compared to the UE conditions.

In the following experiment the *response-signal speed-accuracy tradeoff* (SAT) (Reed, 1973; Dosher, 1976; Wickelgren, 1977; McElree & Dosher, 1989) method was used to test for this possibility. Apart from testing an alternative explanation, this method is attractive in its own right because it provides a detailed picture of the truth-evaluation process. In the predictions above, it was stated that judgment times are ambiguous because they consist of two components, the decision

and the non-decision time. By experimental manipulation of the total processing time in each trial, the following experiment provides more direct access to the different components of the evaluation process than ordinary sentence-picture verification.

5.5 EXPERIMENT 4: RESPONSE-SIGNAL SPEED-ACCURACY TRADEOFF PROCEDURE

In the present experiment, participants first read a question that contained either *more than n* or *fewer than n*. Afterwards, they inspected a picture showing a number of target objects for a short amount of time. Finally, they provided an answer to the question. A response-signal SAT procedure was used in which the time between the onset of the picture presentation and the response was systematically manipulated: After a varying amount of time, a beep tone was played which prompted the participants' responses. After the beep, they had to respond within 300 ms. This time span is so short that they were arguably forced to base their decision on the processing that had taken place before the response signal. Because the amount of processing time was under experimental control, this procedure allows us to test whether the findings from the previous experiment can be explained in terms of a SAT as sketched in the previous section.

5.5.1 *Methods*

5.5.1.1 *Participants*

Forty participants from the University of Tübingen were recruited (29 female, mean age: 24.4 years). They were naïve to the purpose of the study. They received a small financial reward.

5.5.1.2 *Materials*

Example questions that were used in the present experiment are shown in example 5:24. All questions contained either the modifier *mehr* ('more') or *weniger* ('fewer') and one of the numerals *sieben* ('seven'), *acht* ('eight') or *neun* ('nine'). Each question asked about the number of shapes in a picture. Possible shapes were circles, squares, triangles or crosses.

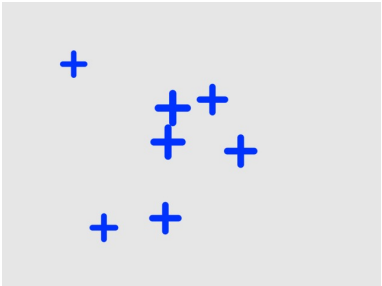


Figure XXXII. Example of a picture used in Experiment 4.

(5:24) a. Sind das mehr als sieben/acht/neun Kreuze?
Are those more than seven/eight/nine crosses?

b. Sind das weniger als sieben/acht/neun Kreuze?
Are those fewer than seven/eight/nine crosses?

Questions were paired with pictures showing randomly distributed, blue objects of the mentioned shape. An example picture is shown in Figure XXXII. In one type of sentence-picture pair, the number of objects matched the numeral in the sentence. In total, 120 pictures of this type were generated – 40 for each of the three numerals. Each of these was paired once with a sentence containing *more than n* and once with a sentence containing *fewer than n*. This resulted in 240 sentence-picture pairs, which were the critical stimuli of the experiment.

In addition to the critical pairs, 960 filler pairs were constructed. The difference between the critical pairs and the fillers was that in the latter the pictures showed a number of objects that deviated from the numeral in the sentence. In particular, each question was paired with four additional pictures that showed $n-2$, $n-1$, $n+1$ or $n+2$ objects of the mentioned shape. In total, 480 such additional pictures were generated (120 per deviance, 40 per deviance and numeral) and each of them was paired once with a question containing *more than n* and once with a question containing *fewer than n*. As a result, half of the fillers required a *no* response and the other half required a *yes* response.

The sentence-picture pairs were presented with one of five response signal lags: 0*ms*, 150*ms*, 350*ms*, 650*ms* or 950*ms*. The lag determined the time frame in which the response had to be given (see next section for details). The sentence-picture pairs were distributed over 10 lists consisting of 600 trials each. In generating the experimental lists, it was ensured that each of the pictures was presented

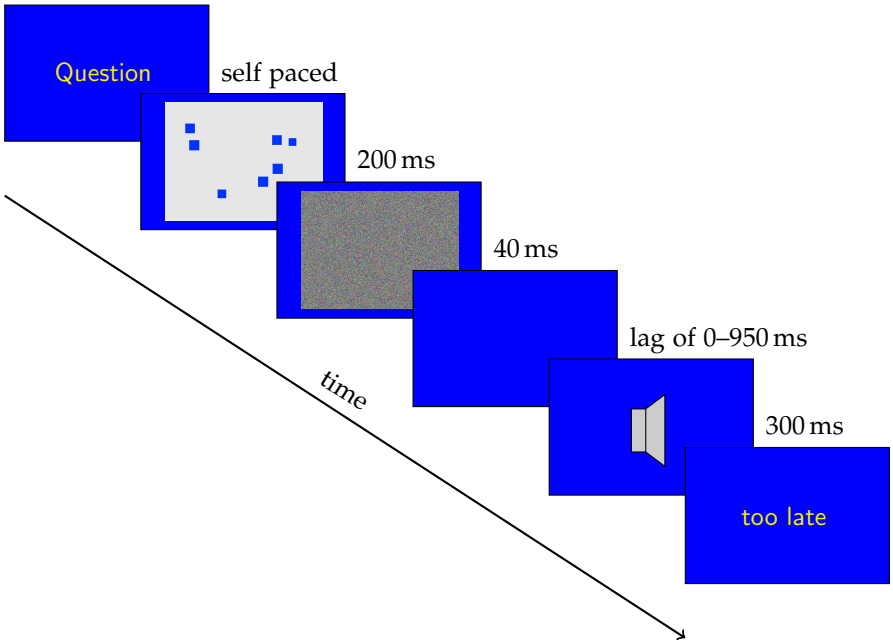


Figure XXXIII. Trial structure of Experiment 4. *Note:* The speaker icon indicates that a 100 ms beep tone was played. It was not shown during the experiment.

only once per list and that each combination of modifiers, numerals and deviances was presented four times per lag, per list. Across each experimental list 60% of the trials required a *no* judgment.

5.5.1.3 *Procedure*

The experimental procedure is graphically sketched in Figure XXXIII. Participants initiated each trial by pressing a button on a keyboard. Next, they read a question self-paced. After reading the question, they pressed a button that initiated the presentation of a picture which lasted for 200 ms. Afterwards, the picture was replaced with image noise. This lasted for approximately 40 ms and was intended to prevent after images. After that, a lag period started, where a blank screen was displayed. The length of the lag period was determined by the experimental condition and ranged between 0 and 950 ms. After the lag period, a beep tone was played for 100 ms. After the onset of the tone, participants had to respond within 300 ms by pressing one of two buttons. If they did not respond in time, the message *too late* was displayed and the next trial started automatically. If they responded too early, the response was not counted.

Assuming no further delays between the individual events in a trial sequence, the total time from the onset of picture presentation until the beep tone ranged between 240 ms and 1190 ms. Depending on the lag, it could be 240 ms, 390 ms, 590 ms, 890 ms or 1190 ms, respectively. Realistically, there was some additional delay, however. For example, misalignment of event timing and the refresh rate of the monitor or loading of images or sounds prior to presentation, may have caused extra delays.

An experimental session took about one hour. At the beginning of the experiment, there was a practice session where participants practiced to respond to the response signal in time. They performed a memory task (the *Sternberg task*) with similar response signal lags as in the main experiment. Before the experiment was started, participants had to respond within the 300 ms time window in 16 out of 15 practice trials. Otherwise, practice was reiterated. The experiment consisted of four blocks. Participants could rest between blocks. At the beginning of each block, two calibration stimuli were presented (cf. Izard & Dehaene, 2008): Written instructions informed participants that a picture of seven or nine objects will be shown to them. After they pressed a button, a picture of seven or nine objects, respectively, was presented for 200 ms.

5.5.1.4 *Predictions*

The IPM predicts that the decision process starts later in questions with *fewer than n* than in questions with *more than n*. The reason for this is that, in the former conditions, an additional operation corresponding to scale-reversal has to take place to provide the representation on which the decision process is based. No such process is assumed in *more than n*. Furthermore, it is predicted that, in the critical trials, where the number of presented objects matches the numeral in the sentence, the accumulation of information during the decision process is faster in question with *fewer than n* than in questions with *more than n*. This prediction was derived from the assumed asymmetric representations of the number of presented objects.

In order to test these predictions, it was planned to calculate d' values (defined in equation 5:25) for the different lags, t , and fit

the function defined in equation 5:26, which approximates how d' develops over time, to these values.

$$d'(t) := Z(hitRate(t)) - Z(falseAlarmRate(t)) \tag{5:25}$$

$$\widehat{d}'(t) := \begin{cases} \lambda\left(1 - e^{-(t-\delta)\beta}\right), & if\ t \geq \delta \\ 0, & otherwise \end{cases} \tag{5:26}$$

The sensitivity index d' is a measure of accuracy adjusted for response bias (see Green & Swets, 1966). In equation 5:25, Z is the inverse cumulative distribution function of the standard normal distribution. The hit rate was calculated as the proportion of correct judgments whereas the false alarm rate was estimated from a suitable control condition that contained the same modifier (see next section). The function in equation 5:26 has three parameters: The parameter λ determines the *asymptotic accuracy*; δ determines the *time intercept* at which d' starts to differ from 0; i.e. chance performance; and β is the *growth rate*, which determines how fast the asymptotic accuracy is approached.

Data obtained in response-signal SAT experiments are usually analyzed this way (cf. McElree & Dosher, 1989). Although there is, in general, no one-to-one correspondence between the parameters of equation 5:26 and the parameters of the DDM (see Ratcliff, 2006), the first prediction mentioned above, namely that the decision process starts later in the DE than the UE cases, can be translated into the prediction that with *fewer than* the time intercept, δ, is larger than in conditions with *more than*. The second prediction, that the rate of accumulation in the critical trials is faster with *fewer than* than with *more than*, may affect the estimated growth rate, β, as well as the asymptotic accuracy, λ. In addition, the effect of the second prediction is sensitive to the control conditions that are used to estimate false alarm rates. It was expected that either of these two parameters would be larger in critical trials with *fewer than* as compared to *more than n* (for details concerning predictions of the DDM with regard to response-signal SAT experiments see, e.g., Bogacz et al., 2006 and Ratcliff, 2006).

The log-Gaussian sequential sampling model amended with negation can be spelled out in alternative ways in order to derive predictions for the present experiment. At least under one of these, the same predictions as just described can be derived. In particular, we may assume that in the DE conditions the decision process continuously

outputs partial information, i.e. the current value of the decision variable, to the negation processing step. The latter simply maps the two response boundaries to the opposite response alternatives. In this scenario, predictions are identical to those of the integrated model. If we assume, on the other hand, a strictly serial organization of processing components where the negation processing step takes place only after the decision process has terminated, predictions are more difficult to derive (suggestions of how such a model can be specified were made by Ratcliff, 2006, although in a different context).

5.5.2 *Results*

Overall, participants responded within the 300 ms interval in 77.7% of the cases. At longer lags, they responded in time more often than at shorter ones. At the five different lags, they responded in time in 51.6%, 72.4%, 85.7%, 89.8% and 90% of the cases, respectively. This was consistent across the two modifiers and the three numerals with a maximal difference below 5% from these means. In almost all combinations of modifiers and numerals, conditions with *fewer than* led to slightly fewer responses in the required time window than their counterparts with *more than* (maximal difference: 5.9%).

Accuracy increased with processing time. Of all trials, 33.7%, 48.6%, 61.1%, 66.5% and 68.0% received a correct judgment at the respective lags. Across all lags, conditions with *fewer than* were answered correctly slightly less often on average than conditions with *more than*. The mean difference in accuracy was 3.9%. Moreover, trials in which the number of presented objects matched the numeral in the sentences were the most difficult ones. At the longest lags, these were judged correctly in 55.6% of the trials. Moreover, trials in which the number of objects was below the numeral in the sentence received fewer correct judgments (deviance of -1: 60.0%, deviance of -2: 76.0% for longest processing times) than trials in which the number of objects was above the numeral (deviance of 1: 70.9%, deviance of 2: 77.8% for longest processing times). Furthermore, conditions with *more than* led to slightly more *yes* responses than conditions with *fewer than*. In the former, 49.9% of all responses were *yes* responses. In the latter, 41% of all response were *yes* responses.

For each combination of numerals, modifiers and lags, d' values were calculated (see equation 5:25). The proportion of correct judgments was used as hit rate. To estimate the false alarm rate, the pro-

portion of incorrect judgments in the easiest condition with the same modifier that required an opposite judgment was used. Estimates of false alarm rates always involved the smallest numeral, *seven*. For instance, the false alarm rate for *fewer than eight* in combination with a deviance of 0 was estimated from the proportion of *no* responses in the condition with *fewer than seven* and a deviance -2.

The d' values for the deviances -2, 0 and 2 are depicted in Figure XXXIV. The predictions for the present experiment were mostly focused on a deviance of 0. In this case, d' values of *more than n* and *fewer than n* were close together at the shortest two lags, but at longer lags, *fewer than* received higher values than *more than*. This pattern is compatible with the predictions. In order to test the predictions, goodness of fit of functions as in equation 5:26 was evaluated.

The following procedure was carried out for each of the three deviances shown in Figure XXXIV. First, the function in equation 5:26 with parameters δ, β and λ was fit to the data using the `nls` function of R (R Core Team, 2016). Then, it was fit separately to the UE and DE cases with three free parameters for each of the two modifiers. Finally, for any subset of the three parameters, one model was fit in which these parameters were constrained to be pairwise identical for the two modifiers. For example, the data were fit assuming different time intercepts and different growth rates but identical asymptotic accuracy for the two modifiers. The model predictions of the latter combination of parameters are shown in Figure XXXIV along the observed d' values. To evaluate goodness of fit, adjusted R^2 values, which take into account the number of free model parameters, were calculated (see McElree & Dosher, 1989). For each of the three deviances, the model with different time intercepts and growth rates but identical asymptotic accuracy was among the models with best model fit. Concerning deviance 0, it was the best model with an adjusted R^2 of .944. For deviance -2, it was on par with the model with different δ and λ but identical β: adjusted $R^2 = .984$. For deviance 2, it was the third best model with an adjusted R^2 of .939 (superseded by the model with different δ and λ but identical β: adjusted $R^2 = .941$; and by the model where only δ differed between the two numerals: adjusted $R^2 = .940$).

In the best models, the time intercept differed between the two modifiers and it was consistently longer in conditions with the modifier *fewer* as compared to the modifier *more*. In the curves depicted in Figure XXXIV, the difference between time intercepts was 305*ms*,

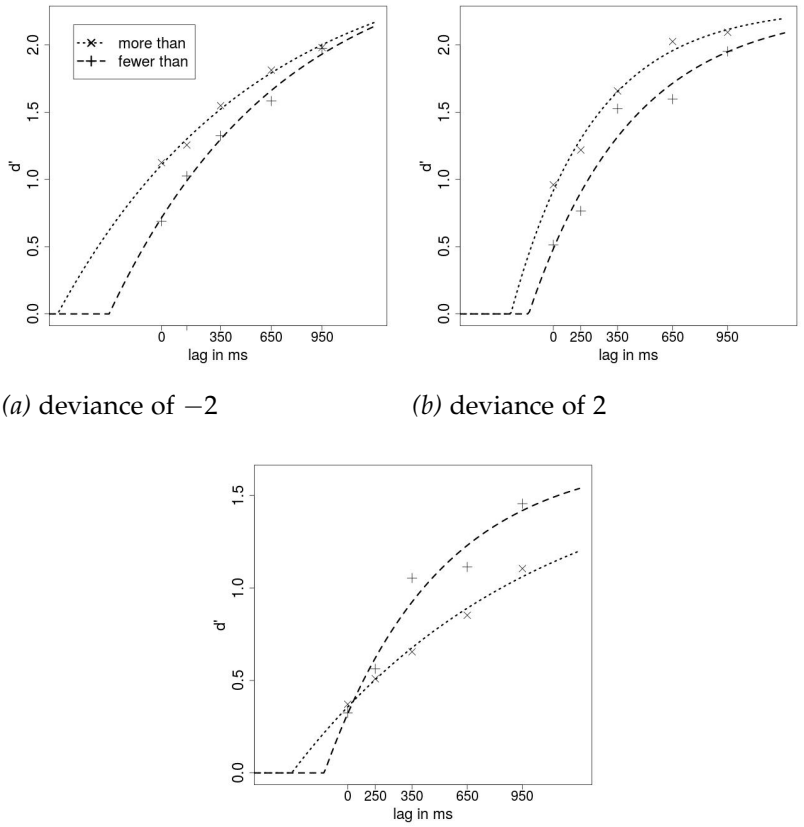
176*ms* and 98*ms* for the deviances −2, 0 and 2, respectively. To get an idea which parameters contribute significantly to model fit, the LRT was used to compare nested models. However, with only ten data points per deviance, these statistics have to be interpreted cautiously. Concerning deviance 0, reducing the best model to one with identical β or with identical δ led to significantly worse model fits ($\chi^2(1) = 12.45, p < .0001$ and $\chi^2(1) = 4.29, p < .05$, respectively). Moreover, the inclusion of separate asymptotes for the two modifiers did not lead to a significant improvement ($\chi^2(1) = 1.24, p = .26$). For deviances 2 and −2, simplifying the best models such that the two modifiers have identical time intercepts reduced model fit significantly ($\chi^2(1) = 12.42, p < .0001$ and $\chi^2(1) = 18.38, p < .0001$, respectively) .

5.5.3 *Discussion*

The motivation for the present experiment was to substantiate the conclusions drawn from the previous experiment and to rule out an alternative explanation based on a SAT. Two qualitative predictions were made. Firstly, it was predicted that, in comparison to sentences with *more than n*, the decision process starts delayed in sentences with *fewer than n*. Secondly, it was predicted that, if the numeral in the sentence matches the number of presented objects, the rate of accumulation is faster for *fewer than n* than it is for *more than n*.

Both predictions were confirmed. Across deviances, the time intercept was estimated to be longer for the modifier *fewer* as compared to *more*. The estimated size of this effect is compatible with what was observed in the previous experiment and it is also compatible with previous studies (see section 5.1). In addition, for a deviance of 0, the growth rate of d' values was estimated to be larger for *fewer than n* than for *more than n*. While the growth rate parameter, β, of the function in 5:26 does not correspond directly to the drift rate in the DDM, it was expected that either β or λ or both would be affected (cf. Ratcliff, 2006).

In the discussion of the previous experiment, an alternative explanation in terms of SAT was discussed. In particular, the findings of the previous experiment could be explained under the assumption that participants take more time for the decision process if they evaluate *fewer than n* than if they evaluate *more than n*. Because the processing time was under experimental control in the present exper-



(a) deviance of -2 *(b)* deviance of 2

(c) deviance of 0

Figure XXXIV. The plots show (1) d' values for *more than n* and *fewer than n* with deviances -2, 0 and 2 (2) curves fitted according to equation 5:26 with different time intercept δ and growth rate β but identical asymptotic accuracy λ for the two modifiers. The legend in panel a applies to all three panels.

iment, this kind of explanation appears implausible in the light of the results.

The present experiment provides further support for the IPM and the results are also compatible with the log-Gaussian sequential sampling model with negation. As noted in the predictions, these two models cannot be distinguished on the basis of the data presented here. As before, the data are hardly compatible with the automata model even if it is amended with an extra processing step. This is because the automata model makes no predictions about the proportions of errors and especially about how they depend on processing time. Furthermore, the data also seem difficult to explain within the log-Gaussian sequential sampling model (Dehaene, 2007) as long as no additional processing step is assumed for *fewer than n* as compared to *more than n*.

5.6 GENERAL DISCUSSION

The present chapter used the processing difficulty of comparative modified numerals in sentence-picture verification as a test case to investigate the relation between formal semantic theory, processing models and experimental data. Recent models and hypotheses from formal semantics, psycholinguistics and cognitive psychology served as background. In addition, an IPM was proposed that combines insights from these areas. In the present section we discuss the theoretical implications that follow from the present chapter and consider how these relate to the bigger picture.

5.6.1 *Theoretical implications*

The experimental results are fully compatible with the IPM and thus support it. Moreover, they are also compatible with a log-Gaussian sequential sampling model (e.g Dehaene, 2007) if it is amended with an additional processing step in the DE case, corresponding, for example, to covert negation. As discussed in the predictions above, these two types of models cannot be distinguished on the basis of our data. However, it was argued in section 5.2.3.1 that the integrated model is *a priori* more plausible because of linguistic considerations. How to accommodate the experimental findings in the other discussed theoretical alternatives is not obvious.

In general, it seems that we need to take into account both the compositional linguistic encoding of truth conditions and the process of numerical comparison in order to model the truth evaluation of modified numerals (see Deschamps et al., 2015 for similar conclusions regarding other UE and DE quantifiers, which we discuss in the next chapter). This is in line with the ITT (Lidz et al., 2011, Hypothesis 4:12). According to this hypothesis, the general cognitive architecture computes, in a transparent way, the operations that are used to specify the truth conditions of a sentence.

As is usually the case, there are conceivable alternative explanations that cannot be excluded ultimately. What the present experimental results and the data discussed in section 5.1 show rather clearly is that there is an additional processing step – or at least a prolonged one – that takes place during the verification or falsification of DE as compared to UE comparative modified numerals. The experimental results are, however, not informative as to what this additional step is. Two potential alternatives are commented on briefly here.

Firstly, we cannot exclude the possibility that pragmatics play a role. Pragmatic accounts of processing difficulty in verification tasks have been discussed in connection with negated sentences and there are some parallels to DE quantifiers. Similar to what is proposed in the IPM, one prominent type of account of processing difficulty during the verification of negated sentences assumes that negation introduces an additional processing step (either in terms of 'proposition matching' or in terms of mental simulations; see e.g. H. Clark & Chase, 1972; Kaup et al., 2007). In contrast, pragmatic accounts propose that negated sentences require specific contextual licensing and cause difficulty if none is provided. Experimental evidence for pragmatic accounts was reported by Nieuwland and Kuperberg (2008) and Tian, Breheny, and Ferguson (2010), for example. One specific pragmatic account claims that difficulty is due to the accommodation of a question under discussion (QUD) if negated sentences are presented out of the blue (Tian, 2014). Moreover, considerations of this kind have recently also been related to the online processing of sentences with DE quantifiers (e.g. Urbach, DeLong, & Kutas, 2015; Nieuwland, 2016). The latter studies show that contextual support can in some but not in all cases annihilate delayed processing of such sentences.

I am not aware of a worked-out pragmatic explanation of the kind of effects discussed in the present chapter. If such an account were to be developed, a number of issues would have to be addressed (see Deschamps et al., 2015 for related discussion). Firstly, one crucial justification for pragmatic accounts of processing difficulty induced by negation does not carry over to DE modified numerals. In particular, negated sentences are, without context, often relatively uninformative (i.e. their *information content* is low; cf. Oaksford & Chater, 2009, who refer to Oaksford & Stenning 1992). This does not generally apply to DE modified numerals. Therefore, alternative justification would have to be given. Secondly, one would have to spell out what the specific licensing conditions of the DE modified numerals are and how they differ from the UE ones. Especially with regard to Experiment 4, in which *yes-no* questions had to be answered, it is far from obvious what a pragmatic account based on implicit QUDs could look like. Moreover, there should be a plausible argument for prevalence of pragmatic effects in experimental tasks where participants are asked to decide about the truth or falsity of sentences over many trials in a repetitive fashion. Furthermore, it should be outlined how pragmatics can account for processing difficulty of sentence-picture verification that shows up during the verification stage, after reading and comprehension are already completed. Finally, it would have to be explained how UE *more than n* may lead to more errors than DE *fewer than n* although it is pragmatically simpler to process.

The second alternative, I want to mention was proposed by Bott et al. (n.d.), who presented a novel perspective on the semantics of quantifiers (see also Bott et al., 2013). Their model is discussed in the next chapter. Therefore, I will not go into details here. Let me just note two things: Firstly, predictions of the Bott et al. model regarding the cases studied in the present chapter depend on additional auxiliary assumptions or linking hypotheses. Secondly, as applies also to a potential pragmatic explanation, the model of Bott et al. may potentially explain the longer RTs we observed in the DE conditions, but it does not explain the high proportions of errors in the UE conditions of Experiment 3a or the slow rate of accumulation in Experiment 4. On the other hand, it is shown in the next chapter that one characteristic feature of Bott et al.'s model follows rather naturally from the one proposed here.

One noteworthy aspect of the present proposal is its parsimony. As discussed above, the essential components of the model were mo-

tivated independently. The semantic building blocks of comparative constructions are well-established in linguistics. The assumed representations of numerical information, the way in which they are manipulated and the model of the decision process were motivated and tested extensively in cognitive psychology and neuroscience. In consequence, the proposal comes essentially for free. This increases its credibility relative to the just mentioned alternatives (*Occam's razor*, see e.g. MacKay, 2003, pp.343–354).

5.6.2 *The bigger picture*

How does the IPM apply to the range of different experimental procedures that were discussed so far? As we will see, short reflection on this question provides us with a link to the automata model and leads to a specific implementation of the mITT (Hypothesis 4:20, Kotek et al., 2015). Although all the discussed experiments investigated comparative modified numerals, it is reasonable to assume that they did in fact investigate a range of different processes. As already mentioned in section 5.2.2, one source of variation between these processes may be how a representation of the number of target objects is derived (cf. Dehaene, 1997; Feigenson et al., 2004). In Experiment 3a, participants could freely choose whether to approximate or to count exactly. Considering the substantial amount of errors in some of the conditions, it seems likely that they relied on approximation at least to some degree. By contrast, participants of Experiment 4 were forced to approximate because visual stimuli were presented for only 200 ms. Moreover, in the experiment of Geurts et al. (2010, exp. 3), numerals were within the *subitizing range* (Kaufman et al., 1949). This allowed for exact representations even in the absence of counting. The data of Szymanik and Zajenkowski (2013), on the other hand, indicate that participants relied on precise counting. Finally, in the self-paced counting study of Koster-Moeller et al. (2008), counting was invited by the experimental procedure.

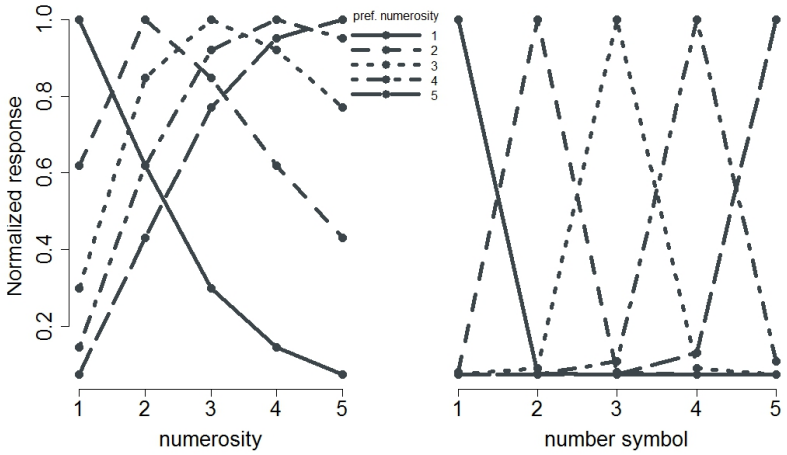
The IPM is applicable to all these cases because it is compatible with different mechanisms for deriving a representation of the number of target objects. When the assumed representations of numbers and numerosities were introduced in sections 5.3.2.2 and 5.3.2.1, respectively, reference was made to the connectionist models of Dehaene and Changeux (1993) and Verguts and Fias (2004). What was not explained in any detail there, is that the network architecture of Verguts

and Fias provides an explicit and plausible model of how to combine exact and approximate number representations. To see how this works, we take a look at their simulation study. In their first simulation, they trained a network that had an architecture as shown in Figure XXXVb to represent numerosities in the range 1 through 5. The input to the *location field* represented objects at different locations.[7] The *summation field* served as a hidden layer that summed up the amount of activity in the location field. Both a supervised and an unsupervised learning procedure led to representations of numerosity as shown on the left hand side of panel a of the figure. This accords with the type of numerosity representation assumed in the present proposal. In a second simulation, the architecture in panel c was used. This was a two-step simulation. In the first step, the network learned to represent numerosity. In the second step, it was trained to also represent symbolic number. Number symbols were modeled as individual nodes of the *symbolic field*. These were presented to the network simultaneously with the corresponding non-symbolic numerosities while it performed an unsupervised learning procedure. After learning, symbolic input was represented as shown on the right hand side of panel a whereas the representation of non-symbolic numerosity was as before.

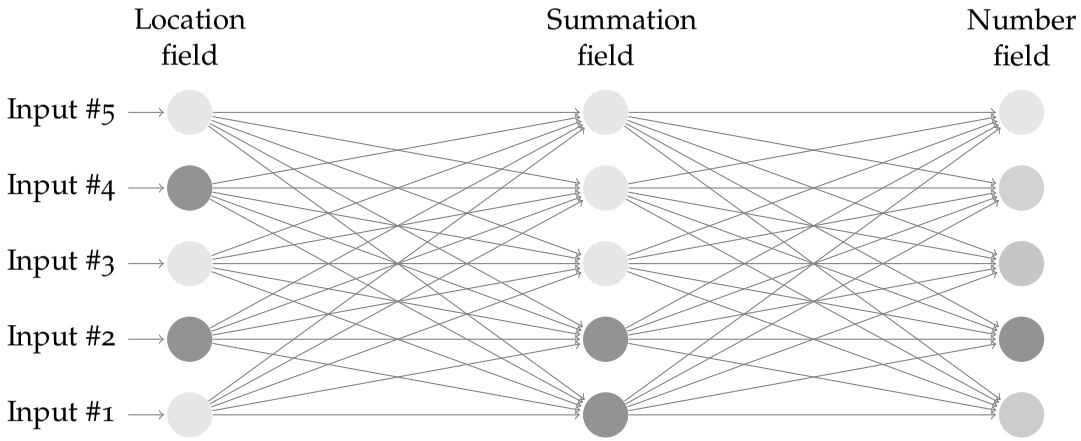
Two points are important here. Firstly, the representations that emerged in the simulations of Verguts and Fias are completely compatible with the IPM. And in particular, they implement Hypothesis 5:18, which states that symbolic number reuses the representation format of the ANS (for further discussion see e.g. Nieder & Dehaene, 2009). Secondly, and more importantly at this point, the model of Verguts and Fias establishes a link to semantic automata. In particular, as far as automata like those shown in Figure XXII are reasonable models of counting routines, it is conceivable that a similar device receives input from the location field and feeds into the symbolic field, cf. Figure XXXVc. Moreover, the resulting representation in the number field can then feed into the IPM, as depicted in Figure XXIX. Incorporating counting procedures along these lines, allows us to apply the model to all the experimental procedures discussed above.

On a more abstract level, we see that, even if verification of quantified sentences proceeds along these lines, semantic automata will still be appropriate tools to reason about the minimally needed computa-

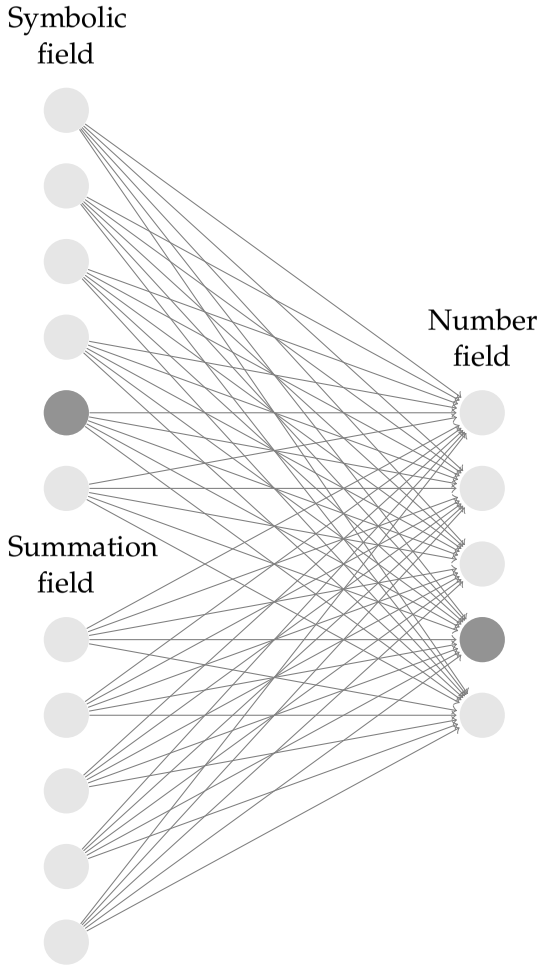
[7] In the original model of Dehaene and Changeux this information was represented in the second layer, after objects on a "retinal image" had been size-normalized.



(a) neural respose



(b) neural architecture



(c) with symbols

Figure XXXV. Sketch of neural networks that model detection of numerosity and number symbols (Verguts & Fias, 2004).

tional recourses of a given verification problem. This is because, in case of flawless performance, the input-output function assumed in the automata model and that of the IPM can be considered identical. Thus, conclusions drawn from the automata model can be considered valid as long as we restrict ourselves to flawless performance.

If extended as just sketched, the IPM opens up the possibility for concrete implementations of the mITT. The most obvious example is the following. If visual stimuli are presented for very short durations the 'symbolic route' is blocked because there is not enough time for counting, but the approximate route is still a viable option. Generally, approximation of numerosity may be a fast and automatic process (cf. Dehaene, 1997; Feigenson et al., 2004; Halberda et al., 2006) whereas counting is slower and only used if appropriate in a given situation. The key point is that one and the same compositional specification of the truth evaluation process can be combined with different ways to derive a representation of how many target objects there are. An interesting consequence of this approach is a natural interface between formal semantic theory and various models of sentence-picture verification.

A comment on the level of analysis (Marr, 1982) may be in order. In developing the processing model, references to results from theoretical and experimental neuroscience were made. The purpose of these was to motivate the representations and computations that are part of the model. However, the present proposal should not be understood as a model at the level of neural computations. Marr argued that cognitive information processing systems must be understood at several distinct but related levels. He famously introduced the three levels of (1) *computational theory*, (2) *representation and algorithmic* and (3) *hardware implementation* (see section 2.1.1). Because the present chapter focused on processing difficulty that was reflected in RTs and proportions of errors, an analysis at Marr's second level naturally suggests itself and the proposed IPM is intended as such. At the same time, care was taken to formulate the model in a way that is consistent with descriptions and analyses at the other two levels. At the computational level, a tight connection to linguistic theory is essential, which – among other things – describes how truth conditions are derived compositionally from the lexical parts of a sentences. In addition, the theory of decisions and inferences on the basis of uncertain information is important at this level. At the level of hardware implementation, neurophysiological studies and computational simu-

lations of how quantitative and numerical information is encoded are essential. Moreover, the computations that are carried out with these representations should be such that they can plausibly be realized at the neuro-computational level. Altogether, the integrated model was formulated such that it may eventually fit into a coherent descriptions of the phenomenon at all three of Marr's levels.

Two important questions are whether the IPM can be scaled to other constructions beside comparative modified numerals and other processes beside sentence-picture verification. These are discussed in chapters 6 and 7, respectively. The answer to the first question is positive, but an interesting follow-up is how far we get with the present approach. The second question seems more difficult. In chapter 7, some ideas are discussed and experimental data are presented that demonstrate enhanced processing load of DE vs. UE quantifiers already during reading.

6

MORE QUANTIFIERS, MORE MODELS: POSSIBLE EXTENSIONS AND RELATION TO OTHER APPROACHES[1]

The previous chapter had a narrow focus on the truth evaluation of comparative modified numerals. The present chapter widens this focus and is concerned with the question whether and how the IPM can be generalized beyond comparative modified numerals. Furthermore, its relation to some other models of quantification is discussed.

For a start, let us put these questions into context. It is no exaggeration to say that "our empirical and mathematical knowledge of quantification in natural language has exploded" (Keenan, 2006, p. 302) since Montague's (1973) pioneering work. Since then "we have witnessed three main stages of research: [g]rand uniformity (the 1970s and 1980s) [...,] [d]iversity (the 1980s and 1990s) [...and] [i]nternal composition (from 2000 on)", as Szabolcsi (2010) explains in her recent research survey on quantification. To her list can be added research on the processing of quantifiers, which mostly took place since the year 2005 (see e.g. the discussion in chapter 4).

Early research was aimed at discovering "semantic universals" that apply to *all* natural language quantifiers in *every* natural language. One well-known example is the *Conservativity Universal* (Barwise & Cooper, 1981; Keenan & Stavi, 1986), which is discussed in the next section (cf. also section 2.2.1). In contrast, more recent research, including that presented in the previous chapters, aims to

[1] A modified version of the model summarized in section 6.3 was published in the proceeding of the Amsterdam Colloquium 2013 Bott, Klein, and Schlotterbeck (2013). Moreover, section 6.3.1 was taken from Bott, Schlotterbeck, and Klein (n.d.) in slightly modified form. These authors which also report the experiments referrred to in section 6.3.2. Parts of the latter work were also presented at the workshop on "Experimental Approaches to Semantics" at the European Summer School of Logic, Language and Information 2015 in Barcelona (Bott, Klein, & Schlotterbeck, 2015), as well as the workshop on "Linguistic and Cognitive Aspects on Quantification" 2015 in Budapest.

obtain detailed pictures of individual quantifiers taking into account rich and diverse empirical data.

A consequence of this development is that the more recent proposals do not attain the level of generality of some of the earlier ones. For example, Hackl (2000), who devolped the semantic fundamentals on which the IPM is based, set out to analyze a range of different comparative quantifiers but eventually conceded that "[f]uture research has to show how the treatment of *more than three* can be extended to cover more complicated comparative determiners like *more than half* as well as amount comparatives like *more books than Bill*" (p. 246). It is remarkable that there is, as far as I know, still no analysis of *more than half* that meets his standards – even more so as both the questions raised by Hackl (2000) and also proportional quantifiers have received quite a bit of attention in subsequent years.

While there are good reasons to shy away from what may be considerd over-generalizations of the early reserach on quantification, it may still be usefull to have at least some kind of roadmap that specifies what the relevant or challenging cases of quantification in natural language are and how one may proceed to investigate them systematically. However, as is explained in some detail below, devising such a roadmap is difficult. All we can do is to focus on the most obvious extensions of the IPM and highlight where difficulties are to be expected in extending the model. As an example, it is shown that the proposal can be straightforwardly extended to comparative proportional quantifiers and existing experimental results are related to this extension. In the final part of the chapter, we discuss the model of Bott et al. (n.d.) and establish a connection to the present one. In particular, it is shown that one of the characteristic features of the Bott et al. model can be naturally derived from – or at least justified by – the IPM.

6.1 WHAT OTHER QUANTIFIERS ARE RELEVANT

The IPM was guided by two hypotheses. The first was the ITT (Hypothesis 4:12) and the second was Hypothesis 5:1, which stated that numerical quantifiers are built upon representations and processes of number cognition (cf. R. Clark & Grossman, 2007). Now, the following question arises naturally:

Question 6:1. *How far do we get modeling quantification in natural language using the building blocks of the integrated processing model?*

To approach this question, it would be useful to know what "quantification in natural language" encompasses or which quantifiers count as relevant. Let us reflect on this briefly. [2] With regard to determiner denotations, which are often considered a subclass of natural language quantifiers, van Benthem (1986) noted the following:

> There are two strategies of description here. One approaches from the outside, so to speak, accumulating global conditions so as to fit to size. The other builds up from the inside, starting from evident cases, and giving an inductive generating procedure. (p.7)

A beautiful example of an approach that combines both of these strategies and demonstrates how they may converge was presented in a classical paper by Keenan and Stavi (1986), who proposed a semantic characterization of the possible interpretations of English "determiner expressions." They started out with a list of examples. Here are a few of them with their original labels:

(6:2)
a. Simplex:
   every, some, two, both, neither, my, your,...
b. Proportional:
   every third, more than 2/3 of the, between 5 and 10% of the,...
c. Modified numerals:
   at least ten, more than two, the ten or more,...
d. Possessives:
   John's, no student's, every teacher's,...
e. Partitives:
   less than five of the ten, the three tallest of the twenty or more,...
f. Boolean combinations:
   not even one, neither John's nor Mary's, most but not all,...
g. a(n)+AP+number of:
   an even number of, a prime number of,...
h. Comparatives with APs:
   more male than female, a prime number of,...

[2] In discussing this question, exact, or almost exact, representations of numerosity are assumed. This allows us to simplify and treat modified numerals in the IPM as truth-conditionally equivalent to modified numerals under their standard GQT based semantics.

i. Comparatives with APs:
more of John's than of Mary's, fewer of the male than of the female,...

...

The determiner expressions in this list share a few obvious properties. For example, they all can combine with what Keenan and Stavi referred to as common noun phrases (CNPs) to form full NPs. Moreover, they all have *extensional* interpretations and they all combine with CNPs that have *countable* denotations. It is evident from the few examples provided here that the class of English determiner expressions has considerable size.

In order to characterize the possible determiner denotations Keenan and Stavi first approached "from the outside." They started by assuming that determiners denote functions from properties to sets of properties (cf. Barwise & Cooper, 1981). Next, they hypothesized that the possible determiner denotations are always *conservative* ("The Conservativity Universal", p. 260, also cf. section 2.2.1) – a hypothesis that is still generally commonly accepted today (for discussion see e.g. Hamm & Zimmermann, 2002). Finally, Keenan and Stavi approached "from the inside." They defined a 'small' set of basic determiner denotations which contained (1) *at least n* ("basic cardinals") and (2) expressions like *Simon's n or more* ("basic possessives"), for every numeral n, and (3) the Aristotelian *every*. They have shown that the closure of this set under, possibly infinite, boolean combinations corresponds exactly to the conservative functions in a given model (see also Hamm, 1989).

What Keenan and Stavi have demonstrated is that it is possible to systematically derive determiner denotations from a small set of 'generators.' They suggested that this fact may be exploited during language acquisition and may substantially simplify the learning problem. Moreover, the general idea that, during learning, determiner denotations are constructed from a small set of primitive building blocks is also implemented in recent statistical learning models (Piantadosi, 2011, ch. 3). [3]

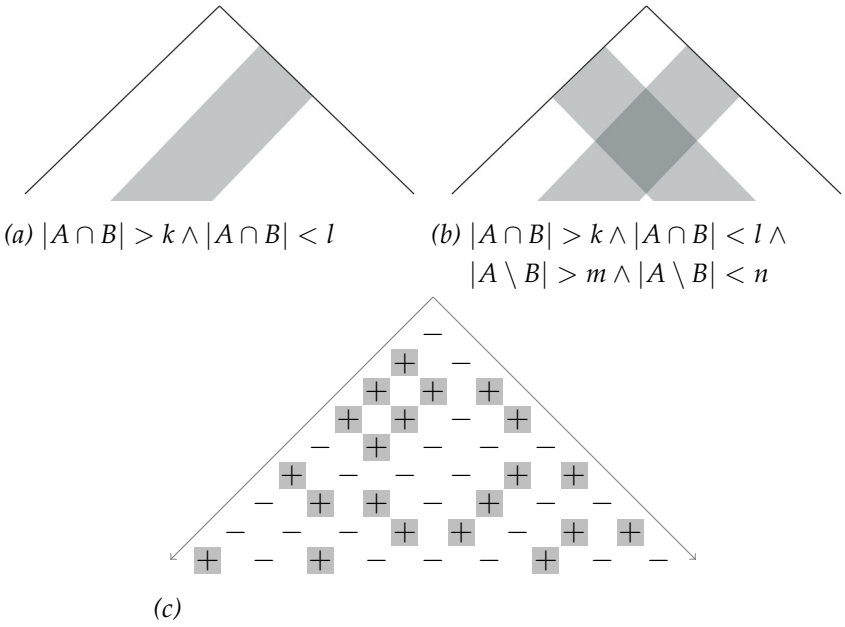
Because the generators of Keenan and Stavi include modified numerals, their result is encouraging and seems to suggest that we can proceed in a similar way to extend the IPM. However, their approach cannot be the model for the present case. This is mainly because of

[3] However, it should be noted here that these authors doubt the usefulness of the Conservativity Universal for learning.

three reasons all of which have to do with the fact that, from our perspective, their set of basic determiners is to small and their set of possible determiner denotations is too large. The first reason is that infinite boolean combinations seem cognitively implausible. Secondly, some relatively simple determiners are derived in implausibly complicated ways (e.g. proportional quantifiers). And finally, the class of determiner denotations singled out by Keenan and Stavi is too large in the sense that it includes cases that are beyond reach from the perspective adopted here.

Let us focus on the last point. What sticks out immediately among the examples in 6:2 are the possessives. It is rather obvious that the building blocks of the IPM do not suffice to cover possessives and this is also not intended. Possessives can be systematically excluded if we restrict determiner denotations to functions that are *isomorphism invariant* – the way we introduced GQs anyway. Moreover, recall from section 2.2.1 that CE-quantifiers can be identified with binary relations between numbers. If we restrict ourselves to these, it may again seem reasonable to systematically derive determiner denotations from the building blocks of the IPM: We could use boolean combinations of modified numerals (and their *inner negations*, cf. Peters & Westerståhl, 2006) to enclose arbitrarily small regions in the number tree (cf. Figure XXXVIa/b). By combining such regions we can in principle define any CE-quantifier. But now we have the same three problems as above. Note in particular that even the CE-quantifiers are far too many to be covered by any cognitively plausible model. It is easy to see that they are uncountably – more precisely $2^{\aleph_0}$ – many. What is more, they contain cases that are clearly not realized in any natural language like, for example, the quantifier that is true of A and B iff either $|A \setminus B|$ is a Fibonacci number or $|A \cap B|$ is a prime number (shown in Figure XXXVIc).

We could continue looking for further conditions to zone in on a set of relevant quantifiers. For example, in terms of computability, we could restrict ourselves to those that are computable by PDA (van Benthem, 1986) or deterministic PDA (Kanazawa, 2013). The former is particularly interesting from our perspective because of its correspondence to expressibility in Presburger arithmetic (recall Theorem 4:8). It is not far-fetched to attempted to enrich the building blocks of the IPM to achieve essentially that expressive power. However, as van Benthem noted himself this class is still way too large. For the time being, it thus seems difficult to find global properties that restrict the



(a) $|A \cap B| > k \wedge |A \cap B| < l$

(b) $|A \cap B| > k \wedge |A \cap B| < l \wedge |A \setminus B| > m \wedge |A \setminus B| < n$

(c)

Figure XXXVI. Number trees again. Panels a/b: defining regions in the number tree via Boolean combinations; panel c: example of an unnatural quantifier

vast space of possibilities to exactly those that are actually realized in natural language.

As a consequence, question 6:1 posed at the beginning of this section cannot get a simple answer. Concerning potential extension of the IPM, we adopt a strategy akin to the second from the quote above. In particular, we shall consider evident cases, study their internal composition and how their building blocks can combine. Despite the non-negligible risk that this may lead to nothing more than a tentative collection of examples (Hamm, 1989, p. 7), the reader is hopefully convinced that this is, at the moment, the most reasonable way forward.

There is a yet another important reason to proceed this way: It has, especially in recent years, repeatedly been argued that "true quantifiers" in natural language are in fact limited to just a few cases and that other "expressions of quantity" cannot be analyzed adequately as GQs (e.g. Nouwen, 2010b; Sternefeld, 2015). The arguments for this claim are not repeated here. They are based on a number of fundamental contributions by Krifka (1999), Hackl (2000), Geurts and Nouwen (2007), among others. In combination with the above considerations, this implies that the class of GQs is on the one hand too large, even if severely restricted, and on the other hand there are

many expressions of quantity in natural language that are not contained within it. The latter form a heterogeneous group that needs to be studied on a case by case basis. This view is clearly expressed in the following quote.

> [T]he GQT notion of a quantifier is not really very suitable if we want to learn more about the semantics of expressions of quantity. If we want to appreciate the subtle differences with which quantifiers communicate quantities, a focus on how they differ is to be preferred over one which sets out to generalise as much as possible. [...] The things we call quantifier are so varied, that they deserve to be studied on a case by case [basis]. (Nouwen, 2010b, p. 254)

This quote contains two rather independent assertions. The first is about the usefulness of GQT to study "expressions of quantity" and only the second is about the heterogeneity of this class and how its members demand individual consideration. I tend to disagree with the first and think that we miss important generalizations and a useful tool for formal analysis if we abandon the GQT perspective. As an example, think of Szymanik's (2010) results about the computational complexity of *polyadic lifts* and how it ironically tells us something general about cumulative readings of numerical quantifiers, which were used by Krifka (1999), among others, as an argument against GQT based analyses. Nevertheless, I agree with the second part of the quote and subscribe to the strategy of case by case analysis for the reasons explained by Nouwen (2010b) and those laid out above.

In the next section, a few rather obvious potential extensions of the IPM are discussed. Boolean combinations of quantifiers are not discussed. We simply note that these are generally possible but, if expressed overtly, usually do not contain more than two parts. How boolean combinations are processed is left open. Hypotheses like the Conservativity Universal are understood as upper bounds. Models of quantification that are able to derive non-conservative determiners are in need of an explanation as to why these never occur. By the same token, models that automatically restrict the possible determiner meanings to conservative ones gain additional credibility. How do the latter points relate to the IPM? In general, the processes and representations involved in number cognition do allow to compute non-conservative quantifiers. Maybe the simplest example is the

Table XXXVII

Class A and Class B modified numerals.

Class A	Class B
more than n	at least n
fewer than n	at most n
under n	up to n
over n	from n
between n and m	from n to m
	maximally n
	minimally n
	n or fewer
	n or more

Note. Taken from Nouwen (2015).

quantifier that is true of two sets A and B iff $|A| > |B|$. However, if we stick to the way meaning is actually assembled in natural language expressions, as dictated by the ITT, we stay within the realm of conservative quantifiers because, by the Conservativity Universal, other quantifiers are simply not realized (cf. the "structural account of conservativity" of Romoli, 2015, who refers to Gennaro Chierchia and Danny Fox).

6.2 POTENTIAL EXTENSION OF THE INTEGRATED MODEL

The most obvious extensions of the IPM are other modified numerals and other comparative quantifiers. Concerning the former, we have to take into account the distinction between what are called *class A* and *class B* modified numerals (Geurts & Nouwen, 2007; Nouwen, 2010a). Examples of both classes are given in Table XXXVII. These two classes are distinguished by several semantic and/or pragmatic properties. The paradigmatic one is their potential for ignorance implications. Class B modified numerals do systematically produce these kinds of inferences whereas class A modified numerals do not. This is exemplified in 6:3. The idea behind this example is that 6:3-b/c are unacceptable because they express ignorance about an obvious truth (cf. Nouwen, 2010a).

(6:3) a. A square has more than three sides.
b. #A square has at least four sides.
c. A square has fewer than five sides.
d. #A square has at most four sides.

Class A and class B modified numerals do not only differ in terms of semantic or pragmatic properties but also with regard to processing and acquisition (Koster-Moeller et al., 2008; Geurts et al., 2010; Bott et al., 2013, n.d.). While class A modified numerals are rather well understood from a semantic stand point, there is an ongoing debate how to analyze class B modified numerals (e.g. Geurts & Nouwen, 2007; Nouwen, 2010b; Schwarz, 2013; Nouwen, 2015; Penka, 2015). Before the basic questions regarding the correct semantic and pragmatic analysis of these modified numerals are answered, an approach that is based on the ITT cannot get off the ground.

Thus, for now, extensions are best focused on class A modified numerals. Specifically, those expressions in Table XXXVII which contain spatial prepositions constitute a potential direction for extension. Constructions like these are found cross-linguistically (for discussion see Corver & Zwarts, 2006). What is interesting in this context, are hypotheses about close relations between the mental representations of space and numerical quantity that have been put forward repeatedly. Specifically, it has been hypothesized that numerical and spatial information is encoded and manipulated using shared representations and mechanisms (e.g. Hubbard, Piazza, Pinel, & Dehaene, 2005; Bueti & Walsh, 2009).

Moreover, the computational model of spatial prepositions proposed by Lipinski et al. (2012) seems relevant in this context. This model was already mentioned when motivating the IPM in section 5.3.2 because the two models show close parallels. In particular, the model of Lipinski et al. is based on a computation that is essentially a generalization of that performed by the comparative morpheme *-er* in the integrated model. It transforms a representation of the spatial position of what is called the *located object* to a representation of its position relative to the *reference object*. The spatial preposition is represented as a region in ('relative') space. The relative position of the *located object* may be within or outside of this region. The latter determines whether a preposition is appropriate in a given context or not. Thus, the preposition plays a role comparable to that of POS in the integrated model.

It is not obvious whether we should expect similar differences in processing difficulty between, e.g., *under n* and *over n* as observed between UE and DE comparative modified numerals. On the one hand, there is no morphologically transparent indication of a scale reversing operator in *under n*. Therefore, we may expect that no such operation

takes place. At the same time, we cannot exclude the possibility that *under n* does involve scale reversal. In particular, we may assume that DE modified numerals always have a derived meaning because the corresponding semantic primitives are lacking. In some cases, e.g. *fewer than n*, the internal composition is transparent, in others, e.g. *under n*, it is not. We cannot decide between these alternatives at this point (cf. H. Clark & Chase, 1972).

Concerning other comparative quantifiers beside comparative modified numerals, we can follow the footsteps of Hackl (2000), and focus on the following types of expressions.

(6:4) a. *Compartive proportional quantifiers*, as in:
Elin read fewer than one third of the books.
b. *Phrasal amount comparatives*, as in:
Elin read more books than magazines.
c. *Clausal amount comparatives*, as in:
Elin read fewer books than Simon shredded magazines.

As regards comparative proportional quantifiers like 6:4-a, extension of the IPM is straightforward if we make one specific assumption about expressions of proportion like *one third*. This is demonstrated in the next section (6.2.1). Moreover, in section 6.2.2, it is discussed how extension to the proportional case relates to empirical data from sentence-picture verification experiments.

Extension to amount comparatives as in 6:4-b/c depends on what syntactic and semantic assumptions we make about the *than*-phrases. This is, however, beyond the scope of the present work. Let me just note that *than*-phrases are often assumed to denote *definite descriptions* of degrees (von Stechow, 1984) and in the case of 6:4-b/c these correspond to numerosity.

Before we discuss comparative proportional quantifiers as an example, Aristotelian quantifiers and *most* are briefly commented on. With regard to the former, I just want to note the following. While these can in principle be analyzed as special kinds of modified numerals (e.g. *some* as *more than 0*), there are several experimental studies that show clear differences between modified numerals and Aristotelian quantifiers in terms of their processing. Sentence-picture verification experiments that show this were summarized in section 4.1.3.2. These data in combination with theoretical considerations (section 4.1.1) give us reason to believe that Aristotelian quantifiers

are subserved by distinct processing mechanisms. Therefore, an extension of the IPM to Aristoitelian quantifiers would not be justified.

The quantifier *most* was already discussed in connection with the automata model and the ITT in chapter 4. The following can be added to this discussion. As an extended version of Pietroski et al.'s (2009) proposal, the IPM is readily applicable to *most*. However, it has to be left open at this point whether extending the model to *most* might resolve some of the conflicts mentioned in chapter 4 regarding the verification procedures associated with this quantifier.

6.2.1 *Comparative proportional quantifiers*

At least if judged from their morphosyntax, comparative proportional quantifiers like *more/fewer than half* are among the closest relatives of comparative modified numerals. Superficially, there are only two differences (but see section 4.1 for some semantic differences). The first lies in how the *than*-phrase is realized: In one case it contains a numeral, in the other it contains *half*. The second difference is that the proportional variant is always a partitive whereas this is optional in the modified numeral variant.

In the present section, the IPM is extended to comparative proportional quantifiers and, in particular, to the following German examples. While the discussion is focused on these examples, it applies to other comparative proportional quantifiers as well.

(6:5) a. Mehr als die Hälfte der Karten sind rot.
More than the half of the cards are red.
'More than half of the cards are red'.

b. Weniger als die Hälfte der Karten sind rot.
Fewer than the half of the cards are red.
'Fewer than half of the cards are red'.

In order to derive symbolic meaning representation of the quantifiers in 6:5, the expression *die Hälfte* has to be integrated into the sentence. The following is the simplest way to achieve this I can think of. We simply assume that expressions like *half*, *a third*, *a fourth*, etc., which we refer to as *expressions of proportion*, denote proportions (cf. Solt, 2016b). Proportions can be conceived of as degrees that are associated with what Kennedy and McNally (2005a) call a *totally closed* scale. In fact, the ability of gradable expressions to combine with expressions of proportion is used by Kennedy and McNally (2005a) as a diagnos-

tic for this kind of scale. Therefore, expressions of proportion can be straightforwardly integrated into the sentence meaning as long as we assume that an operator MANYPROP is available as an alternative to MANY. Of course, MANYPROP has the same syntactic category as MANY. But it maps two properties of pluralities, say **card** and **red**, to a representation of the following proportion (under the obvious interpretation function, $[\![\cdot]\!]$):

$$\frac{[\![\textsc{max}\big(\lambda d.\exists x\big(\textbf{cards}\ x \wedge \textbf{red}\ x \wedge \#(x) = d\big)\big)]\!]}{[\![\textsc{max}\big(\lambda d.\exists x\big(\textbf{cards}\ x \wedge \#(x) = d\big)\big)]\!]} \tag{6:6}$$

The symbolic meaning representation of the quantifier in 6:5-b is derived as shown in Figure XXXVIII. Apart from the mentioned changes, the derivation is parallel to that of the modified numerals discussed in the previous chapter.

I would like to mention three pieces of evidence that support the assumptions just introduced. Firstly, expressions of proportion by and large occur in the same syntactic positions as other degree denoting expressions like, e.g. numerals, measure phrases or intensifiers. Secondly, the existence of a silent operator MANYPROP is plausible in light of the fact that overt *many* also displays an ambiguity between a cardinal and a proportional interpretation (Partee, 1989) and that, thirdly, proportional readings also surface in comparatives, as the following example shows.

(6:7) More residents of Ithaca than NYC know their neighbors.
(from Solt, 2016b, who refers to Barbara Partee)

As was reviewed by Jacob, Vallentin, and Nieder (2012), there is accumulating evidence that proportions are represented using exactly the same format as representations of numerosity. Thus, the symbolic representations can be mapped to essentially the same truth evaluation process as before. What is not discussed here are the implementation of MANYPROP and potential distinctions between approximate and exact representations of proportions.

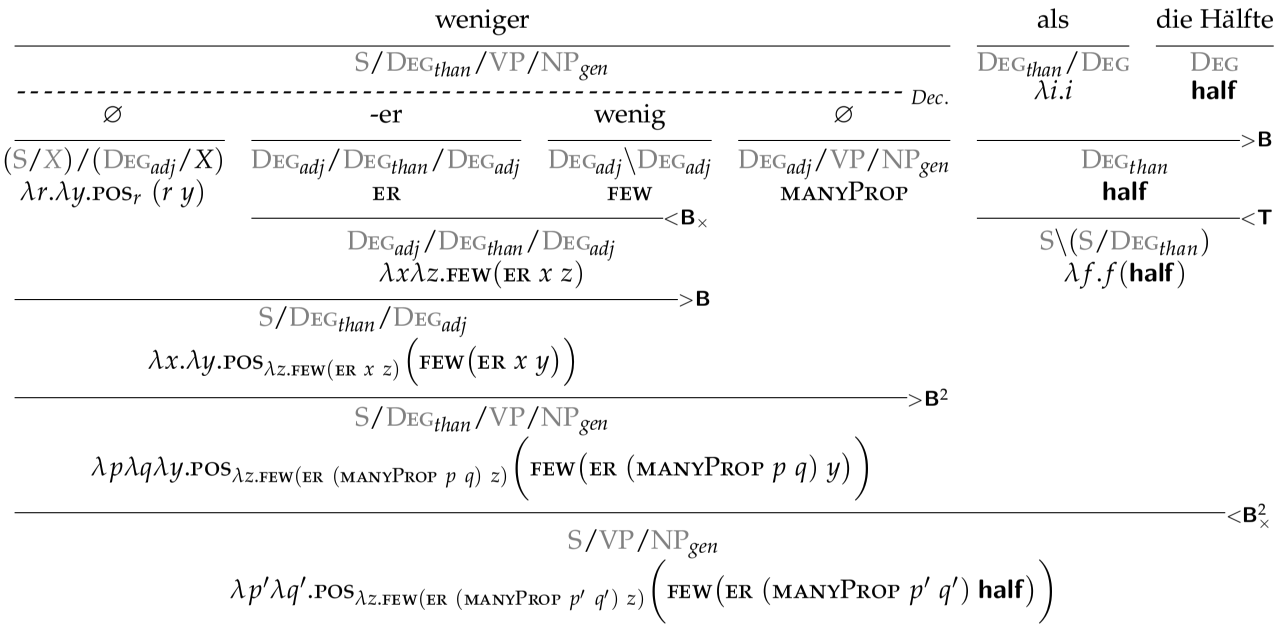


Figure XXXVIII. Derivation of *weniger als die Hälfte* ('fewer than half').

6.2.2 *Existing data from sentence-picture verification*

What does the extension to proportional quantifiers predict and how does it relate to existing data form sentence-picture verification experiments? These questions are discussed in the present section. In chapter 4, a number of relevant sentence-picture verification experiments were discussed. The central finding was that, as predicted by the automata model, comparative proportional quantifiers led to enhanced working memory load during verification as compared to, e.g. Aristotelian quantifiers or modified numerals.

The following difference between modified numerals and proportional quantifiers carries over to the present approach. The truth or falsity of modified numerals depends only on objects that have both the property expressed by the restrictor argument and that expressed by the scope. In contrast, proportional quantifiers depend on all objects that have the restrictor property. It is rather obvious that the latter may cause enhanced working memory involvement. Apart from informal considerations of this sort, predictions about which cognitive resources or taxed specifically by the comparative proportional quantifiers depend entirely on how the operator MANYPROP is implemented. While it is easily conceivable that the corresponding computation requires a substantial amount of working memory, further characterization of this kind of computation is beyond the scope of the present work. What is important to note here is that, once MANYPROP has computed a representation of a proportion, there is no substantial difference between comparative proportional quantifiers and comparative modified numerals anymore.

A direct prediction is that the difference in processing difficulty between UE and DE modified numerals should carry over to UE vs. DE comparative proportional quantifiers. This is because the extra processing step of scale reversal induced by FEW is predicted to take place in the DE versions of both types of quantifiers. I am aware of two existing studies that already tested this prediction. The first is that of Szymanik and Zajenkowski (2013) that was already discussed above in section 5.1. The second was reported by Deschamps et al. (2015). These two studies did, however, report conflicting results. Both are summarized briefly here. Afterwards, it is discussed what may have caused the divergence.

When discussing the study of Szymanik and Zajenkowski (2013) in section 5.1, we focused on the conditions with modified numerals.

In addition to these, Polish versions of sentences like *more/fewer than half of the cars are black* were included in the experiment. Szymanik and Zajenkowski found an interaction between the direction of entailment and the type of quantifier in RTs. DE modified numerals took longer to judge than UE ones but there was no difference between the DE and UE proportional quantifiers.

These results stand in contrast to what would have been expected under the IPM and also to what Deschamps et al. (2015) found. These authors also investigated effects of the direction of entailment (or polarity) of quantifiers on RTs and proportions of errors in sentence-picture verification experiments. They compared the a- and b-variants of 6:8–6:10 and found prolonged response times as well as increased proportions of errors in the b-variants, which contained DE quantifiers. This led to robust effects of the direction of entailment across quantifier types. The DE conditions took about 100–200 ms longer to judge than the UE conditions.

(6:8) a. More than half of the dots are blue.
b. Less than half of the dots are yellow.

(6:9) a. Many of the dots are blue.
b. Few of the dots are yellow.

(6:10) a. More dots are blue than yellow.
b. Fewer dots are yellow than blue.

Like much of chapter 5, their study also investigated the interplay between the direction of entailment of quantifiers and the processing of numbers and numerosities. [4] In addition to polarity, the proportion of yellow to blue dots shown on the pictures was manipulated. The latter manipulation also affected processing difficulty, but there was no interaction with the direction of entailment. The authors concluded that the two experimental manipulations affect different processing components. The polarity manipulation affects linguistic encoding and processing whereas the manipulation of proportions affects numerical processing. Additional support for this conclusion was obtained from non-linguistic control conditions which had in-

[4] A comment on the time line: The study of Deschamps et al. was available online beginning of July 2015. In March 2015, a compact version of the integrated processing model and the results of Experiment 3 as well as design and prediction of Experiment 4 were submitted to the "Experimental Approaches to Semantics Workshop" at ESSLLI 2015 (Schlotterbeck, 2015). At that time I had no knowledge of the study of Deschamps et al. (2015).

structions as in 6:11. In these conditions, no difference between the a- and b-variants were found (cf. Cummins & Katsos, 2010).

(6:11) a.  /
b.  /

Deschamps et al. related their results to recent proposals from the linguistic literature that posit additional compositional structure and thus more complex representations in sentences with DE as compared to UE quantifiers. They call this a "syntactic account" and dismiss several alternative explanations. Their syntactic account is motivated by split scope data in sentences containing downward entailing quantifiers (e.g Bech, 1955; Jacobs, 1982; Penka & Stechow, 2001). Their conclusions are strikingly similar to those presented in chapter 5. A superficial difference between the IPM and the syntactic account of Deschamps et al. is that the latter attributes the difficulty in the DE conditions to syntactic dislocation, i.e. movement operations, whereas the present account only assumes an additional semantic operation but remains silent about movement operations. Apart from this difference, the two proposals are compatible. While not attempted here, it should be possible to apply the IPM to the other quantifiers tested by Deschamps et al. (2015), beside comparative proportional quantifiers, namely the sentences in 6:9 and 6:10. Amount comparatives as in 6:10 were briefly commented on above. A detailed semantic analysis of *many*-sentences like 6:9 was conducted by Solt (2009, 2014). See also references therein and Schöller and Franke (2015, 2016).

What sets the IPM apart, is that it specifies an explicit processing model. As a consequence, explicit predictions about which processing components are affected by experimental manipulations can be derived. In particular, it is predicted that the direction of entailment affects non-decision times whereas the difficulty of the numerical comparison affects drift rates. These predictions can be tested if data are collected in a way that allows for estimation of the DDM parameters (cf. Dehaene, 2007). In order to collect enough data per participant, this would probably require testing participants in multiple sessions.

One aspect of the data reported by Deschamps et al. seems at odds with the IPM at first sight: DE quantifiers led to more errors than UE ones. In an experimental procedure without time limit the extra processing step in the DE version should lead to slowdown, but there

should not be an effect on proportions of errors in the experimental conditions used by Deschamps et al. However, they used a time limit of 3 s. This may have caused the effect. Equal proportions of errors are only expected if the decision process is allowed to reach one of the response boundaries. Otherwise, proportions of errors may be affected in addition to RTs (cf. Experiment 4).

What may be the reason that Szymanik and Zajenkowski and Deschamps et al. obtained conflicting results? There are several possibilities. Firstly, there is the remote possibility that the effect reported by Deschamps et al. is actually due to a confound. Their experiments were well designed, for sure. But there was one confounding factor that is difficult to control for: In their design, there is a mismatch between the dominant color shown on the picture and the color mentioned in the sentence in the true DE conditions. In the false DE conditions, on the other hand, there is a match. This may in principle lead to some kind of interference effect between lexical fit and truth values (cf. Urbach & Kutas, 2010). However, if this explanation was on the right track, one would expect the effect of the direction of entailment to be stronger at the extremes, with very many or very few objects of one color. But this is not what Deschamps et al. found.

Secondly, the discrepancy may also be due to differences in task demands. Whereas approximation did suffice in the experiments of Deschamps et al., proportions were relatively close and precise judgments were required in the experiment of Szymanik and Zajenkowski. This may have affected the verification procedure used by the participants. In particular, it is possible that, in the experiment of Szymanik and Zajenkowski, participants used the kind of "vote counting" or "one-to-one-plus" strategy that was discussed in section 4.2. Under the IPM, this is not expected but would also not be impossible. For example, in an experimental setting in which the canonical verification procedure does simply not succeed, it would be compatible with the model if people switched to a different verification procedure that is effective in that particular task.

However, the most likely explanation of the divergence is that statistical power was low in the study of Szymanik and Zajenkowski and they did therefore not discover the effect. The mean RTs they report for the proportional conditions are over 6 s with standard deviations of around 2 s. If we assume that the true effect of the direction of entailment is around 100–200 ms as was observed by Deschamps et al. and is also compatible with experiments 3a & 4 reported in the

previous chapter, then it is rather likely that the effect was missed in the experiment of Szymanik and Zajenkowski.

6.3 EMPTY-SET EFFECTS IN DIFFERENT KINDS OF DE QUANTIFIERS

Another explanation of processing difficulty with DE quantifiers was proposed by Bott, Schlotterbeck, and Klein (n.d.), who presented a novel perspective on the semantics of quantifiers (see also Bott, Klein, & Schlotterbeck, 2013). They developed an algorithmic theory of quantifier interpretation roughly in the sense of Moschovakis (1994): The *sense* of an expression is the algorithm which computes its *denotation*. They modeled quantificational complexity at two levels. Firstly, the algorithms corresponding to two quantified statements can inherently differ in complexity, that is, one algorithm can be inherently more complex than the other. They take this kind of complexity to affect the difficulty of computing a semantic representation during the comprehension of quantified statements presented out of context. Secondly, the execution of a given algorithm can vary with respect to the number and kind of steps required in order to compute the semantic value. This corresponds to the complexity of a given instance of the verification problem.

One interesting aspect of their proposal is that it automatically restricts determiners denotations to conservative quantifiers. What sets their proposal apart from the semantic automata model, for example, is the assumption of a general asymmetry in processing *positive* vs. *negative instances of a predicate*, i.e. entities that are vs. are not in its denotation. On the basis of this assumption, they predicted that so-called empty-set quantifiers, i.e. quantifiers that have the empty set as a *witness set*, call for a more complex verification algorithm than other quantifiers. [5] In the following, the basic idea of the Bott et al. model is sketched briefly. After that, some of their experimental results are summarized. Finally, it is speculated what connections may exist to the IPM.

[5]

Definition 6:12 (Witness set, Barwise & Cooper, 1981; Peters & Westerståhl, 2006)**.** *Let Q be a CE GQ, A any set and let Q^A of type $\langle 1 \rangle$ be defined, for all M and $B \subseteq M$, by: $(Q^A)_M(B) \Leftrightarrow Q_{M \cup A}(A, B)$. Then we call any set $X \subseteq A$ for which $(Q^A)_M(X)$ is true a witness set of Q^A.*

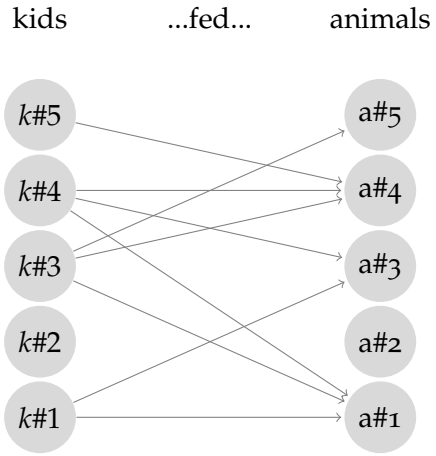


Figure XXXIX. A hypothetical situation in the zoo.

6.3.1 *The model*

Only the basic idea behind the model of Bott et al. is sketched here. Moreover, the discussion is limited to sentences with at most two quantifiers although their model also applies to sentences with arbitrarily many. Consider sentence 6:13 under its surface scope reading (i.e. *more than half of the kids* takes wide scope) to be evaluated in the situation depicted in XXXIX. The picture shows sets of boys and animals as well as the relation denoted by *fed*. Obviously, the sentence is true in this situation because three out of five kids are such that each of them fed either two or three – ergo more than one – animals.

(6:13) More than half of the kids fed more than one animal.

How could an algorithm for evaluating this sentence look like? Bott et al. proposed the following rather simple verification procedure: A doubly quantified sentence with quantifiers Q_1 and Q_2 under the scope reading where Q_1 takes scope over Q_2 is evaluated by successively adding quantifiers to the verb denotation starting with the narrow-scope quantifier. The algorithm iterates through tuples σ from the predicate denotation P. It is described in 6:14. In order to keep the description concise, the notation $\sigma^{[n/x]}$ is used. This simply denotes the tuple σ' that is identical to σ except that the n-th element is replaced with x.

(6:14) **Simple expansion rule** (for a quantifier Q_n, where $n \in \{1,2\}$) For each tuple $\sigma \in P$ collect the set $\{x : \sigma^{[n/x]} \in P\}$. Next, intersect this set with the restriction of Q_n and check whether

the result is a *witness set* of Q_n. If so, add $\sigma^{[n/Q_n]}$ to the predicate denotation.

In the above example, we would start by expanding the second argument slot of *fed* with *more than one animal*. In order to do so, we consider each pair in the denotation of *fed* and check whether the number of animals that the kid agent in that pair fed is larger than one. For example, we could start by considering the pair $\langle k\#1, a\#1\rangle$. Then, we would have to check all pairs that have $k\#1$ in the first position. These are $\langle k\#1, a\#1\rangle$ and $\langle k\#1, a\#3\rangle$. Thus, we have to evaluate the cardinality of the set $\{a\#1, a\#3\}$. Since this cardinality is two, we are licensed to expand the verbal predicate with $\langle k\#1, \textit{more than one animal}\rangle$. Analogously, by application of the simple expansion rule to the other pairs we can add $\langle k\#3, \textit{more than one animal}\rangle$ and $\langle k\#4, \textit{more than one animal}\rangle$. We refer to the set of pairs that are to be added as the expansion set.

It is a distinguishing feature of the simple expansion rule that it often allows us to ignore entities not participating in the relation expressed by the verb. That is, the entities $a\#2$ and $k\#2$ do not play any role for simple expansion. The only exception is if they are needed in order to check whether some set is a witness set of one of the quantifiers involved. The latter is not relevant for the quantifier *more than one animal* but will be when we apply simple expansion to the proportional quantifier, Q_1, *more than half of the kids* . What about the pair $\langle k\#5, a\#4\rangle$? Kid $k\#5$ fed only a single animal, $a\#4$. Therefore, simple expansion does not succeed since the singleton set $\{a\#4\}$ is not among the witness sets of *more than two animals*. We are not licensed to add $\langle k\#5, \textit{more than one animal}\rangle$.

We start the evaluation of the next quantifier, *more than half of the kids*, with the expansion set just computed:

$$\{\, \langle k\#1, \textit{more than one animal}\rangle\,,\\ \langle k\#3, \textit{more than one animal}\rangle\,,\\ \langle k\#4, \textit{more than one animal}\rangle\,\}.$$

Now, simple expansion is straightforward. We just have to check whether the set $\{k\#1, k\#3, k\#4\}$, which consists of three kids that fed more than one animal, is among the witness sets of *more than half of the kids* in the scenario in Figure XXXIX. In order to do so, we need to know the cardinality of the set of kids. Note that at this point, which we may call *witness set identification*, the whole restrictor set

becomes relevant, no matter whether its elements participate in the verbal relation or not. Since the three kids out of a total of five kids is among the witness sets of *more than half of the kids*, simple expansion of *more than half of the kids* yields:

$$\{\langle \textit{more than half of the kids}, \textit{more than one animal}\rangle\}. \qquad (6{:}15)$$

We see that successful expansion with both quantifiers corresponds to verification of the sentence in the given situation. If, however, the pair in 6:15 could not be added, the sentence would be falsified.

To summarize, the simple expansion algorithm consists of a single rule, which is used to add tuples containing quantifiers in addition to individuals to the verbal predicate. The execution of this rule works in the absence of negative predicate instances, i.e., once we know the cardinalities of the restrictor sets (cf. *more than half*) we can safely ignore those individuals not in the relation.

Next, it is demonstrated that this simple expansion procedure may fail if one of the quantified expressions has the empty set among its witness sets (called *empty-set quantifiers*). This was not the case in sentence 6:13. However, if we replace *more than one animal* by *fewer than two animals*, a quantifier that is true in case there are no target objects, simple expansion does not work anymore.

(6:16) More than half of the kids fed fewer than three animals.

The problem is that, for this sentence, not only those individuals are crucial that participate in the relation, but also those that do not. In the situation shown in Figure XXXIX, three out of a total of five kids, namely *k*#1, *k*#2 and *k*#5, fed fewer than three animals. Hence, the sentence is false. However, it is predicted to be true if we apply simple expansion: Expansion with *fewer than three animals* allows us to add the following pairs to the predicate.

$$\{\ \langle k\#1, \textit{fewer than three animals}\rangle\ ,\\ \langle k\#5, \textit{fewer than three animals}\rangle\}$$

Expansion with Q_2 reveals that the simple expansion rule misses one crucial aspect of the situation in Figure XXXIX, namely that $k\#_2$

fed no – ergo fewer than three – animals. As a result, no further expansion is possible. This is undesired because it wrongly predicts the sentence to be false.

This is not an isolated case but a general shortcoming of the simple expansion rule. Bott et al. prove that sentences that do not contain any empty-set quantifiers can be evaluated with the simple expansion rule. However, sentences with empty-set quantifiers, such as *fewer than three animals*, call for a more complex expansion algorithm – an algorithm that allows the interpretation system to not only keep track of positive predicate instances but also of negative ones. They develop such an algorithm in two steps. Firstly, *negative predicate instances* are encoded in addition to *positive* ones. For the example at hand, this means that the Cartesian product of the two restrictor sets is partitioned into two subsets – the positive and the negative predicate instances. Secondly, another rule that positively expands an empty-set quantifier in an empty-set situation is added. With regard to quantifiers that do not have the empty set among their witness sets they propose that the simple expansion operation is used, but for the others the extended *complex expansion operation* is used.

We do not go into any detail here but simply note that the simplest way to think of the additional rule is presumably as an inference from a sentence where the to-be-expanded quantifier is replaced with *no*. For example, with regard to 6:16, this rule would implement the following inference.

(6:17) *k*#2 fed no animals.
∴ *k*#2 fed fewer than three animals.

This inference would in turn license addition of the pair

$$\langle k\#2, \textit{fewer than three animals} \rangle$$

to the denotation of fed.

6.3.2 *Predictions and data*

Since the complex expansion operation is more complex than the simple expansion operation, the theory of Bott et al. immediately predicts processing differences between *empty-set* and *non-empty-set quantifiers*. Furthermore, because all DE quantifiers are empty-set quantifiers the theory explains some processing differences between monotone de-

creasing and other types of quantifiers. Specifically, Bott et al. discuss three predictions of their model. The first is that empty-set quantifiers are especially difficult to process in situations where no target objects are present (called *empty-set situations*). This prediction is derived from the assumption that it is particularly difficult to draw a positive conclusion from the absence of predicate instances. The second prediction is that empty-set quantifiers cause difficulty during comprehension because a more complex verification procedure has to be prepared than in the case of non-empty-set quantifiers. It could, for example, be costly, to retrieve the additional rules *from* memory or to retain them *in* memory. The third prediction is that evaluation of empty-set quantifiers is generally more difficult than evaluation of non-empty-set quantifiers. This prediction depends on whether we assume that the complex expansion operation is used to evaluate empty-set quantifiers in cases where the simple expansion operation would do as well. If we assume that this is the case, we may, for example, expect difficulty because attention is directed to negative in addition to positive predicate instances or because there is a larger set of rules in complex than in simple expansion from which the appropriate has to be chosen.

Bott et al. report evidence from three experiments that supports their model. I would like to focus on the first prediction here, in particular with regard to modified numerals. The reason is that we did not derive this prediction from the IPM, but it is shown in the next section that the IPM can be taken to justify the distinction between simple and complex expansion theoretically.

In all three experiments of Bott et al., participants performed a truth-value judgment task after they had read a sentence self-paced. Evidence for enhanced difficulty of empty-set as compared to non-empty set quantifiers was obtained across experiments. This difficulty was especially pronounced in empty-set situations. For example, sentences like 6:18-a led to substantially more errors and longer RTs than sentences like 6:18-b when evaluated against visual contexts in which all the squares had a different color than pink.

(6:18) a. More than five squares are pink.
b. Fewer than five squares are pink.

This "empty-set effect" manifested itself most clearly in the proportions of errors. In empty-set situations, 6:18-b led to 25% errors whereas there were below 10% errors in all other conditions. Simi-

lar effects were also observed with boolean combinations that were non-monotone (i.e. neither UE nor DE) but had the empty set among their witness sets and also in doubly quantified sentences. The authors rule out several alternative explanations of these effects.

6.3.3 *Relation to the integrated processing model*

There are three questions about the relation between the study of Bott et al. (n.d.) and the IPM. The first is whether the model of Bott et al. can account for the data presented in the previous chapter. The second question is whether empty-set effects, as observed by Bott et al., can be accounted for in the IPM. The third question is whether the two models are compatible.

With regard to the first question, we note that the simple expansion operation would have sufficed to solve the tasks of experiments 3a and 4, even for the DE conditions, because none of the experimental conditions involved empty-set situations. However, participants did not know this in advance because what rules are needed depends on the pictures that are actually presented in each trial. In order to be able to deal effectively with all possibilities, the complex expansion operation would have had to be prepared during reading. The observed increase in RTs of *fewer than n* as compared to *more than n* can be accounted for if we assume that a larger set of rules from which the appropriate one has to be chosen incurs processing cost.

However, the basic ingredients of the IPM are needed anyway and thus the model of Bott et al. can hardly be considered a genuine alternative: Firstly, the result that the UE conditions led to a higher proportion of errors than the DE ones in experiments 3a and 4 is completely unexpected under the model of Bott et al. Obviously, it is possible to account for this effect if the processing of numerical information is taken into account, as is done in the IPM. Secondly, as discussed in section 5.2.3.2, there are also independent reasons to assume some kind of antonym operator to be part of the lexical semantics of *fewer than*.

While the model of Bott et al. is intended to apply to all natural language quantifiers, the authors note themselves that the distinction between simple and complex expansion is not the only source of quantificational complexity. Among other things, they explicitly mention the possibility that the DE comparative modified numerals of the form *fewer than n* may introduce inherent processing difficulty

as compared to their UE counterparts due to their internal composition.

The second question is whether empty-set effects, as observed by Bott et al., can be accounted for in the IPM. While these effects do not follow directly from the IPM, there is one close parallel between the two models. In particular, just as the simple expansion operation, the IPM does also not suffice in empty-set situation. Remember how a representation of numerosity was derived in the IPM. It was assumed that approximate representations are derived as described in the connectionist models of Dehaene and Changeux (1993) and Verguts and Fias (2004). These network models received 'visual' input, specifically the input was a map of object locations. The amount of activity in this map was summed up to compute a "summation code." However, if no target objects are present, then no activation can be summed up and, as a result, no representation of numerosity can be computed.

Moreover, it was assumed that precise number is learned via association of numerosity representations and number symbols (Verguts & Fias, 2004). Again, no association can be learned without any input to the model and, as a result, no precise representations can be computed in empty-set situations either. The claim here is not that people have no knowledge or representation of the number 0. To the contrary, it was assumed in the IPM that the comparative morpheme *-er* semantically corresponds to a kind of subtraction operation and this may of course lead to a representation of 0 and even of negative numbers. What is claimed here instead is that the 'usual' way (or ways) in which a representation of the number of target objects is computed from visual input is blocked or does not succeed in empty-set situations. We do not have to stipulate this, but it can be derived from the IPM.

As a consequence, it is expected that, in empty-set situations, the IPM does not suffice to perform the verification task. However, this applies equally to both UE and DE modified numerals. In order to explain the effects reported by Bott et al., we have to make additional assumptions. In particular, we may assume that the IPM defaults to the "no false" response, in case no representation of the number or numerosity of target objects is computed. Furthermore, we may assume that in the case of *fewer than n* an inference as in 6:17 may be used to override this default. However, such inferences are costly and do not succeed always. But this is the essence of what is implemented

in the model of Bott et al.. From this perspective, the IPM may be taken to justify what is implemented in the model of Bott et al.

Our third question was whether the two models are compatible. As has hopefully become clear from the previous paragraphs neither of the two models can account for all of the data on its own. Rather the model of Bott et al. may be viewed as a description of part of a control structure that manages different potential verification procedures. One of these is what I would call the canonical procedure, as implemented in the IPM for comparative modified numerals. Another potential procedure consists of inferences rules like exemplified in 6:17. If none of the available procedures succeeds the sentence can not be verified and the system defaults to falsification. Under this perspective the two models complement each other.

7

MORE PROCESSES: COMPREHENSION

Two questions are discussed in the present chapter. The first is whether the more complex internal composition of *fewer than* as compared to *more than* that is assumed in the IPM also leads to difficulty during online comprehension, in addition to verification. The previous two chapters were concerned with processing difficulty that emerged while evaluating the truth of a sentence against a picture after the sentence has already been read and understood. These data are complemented in the present chapter, which focuses on difficulty that emerges already during reading, before a truth-value judgment is performed. Increased difficulty of *fewer than* vs. *more than* during reading is demonstrated using eye tracking data recorded while subjects read German sentences containing these modifiers. The second question that is discussed in the present chapter is whether and how the IPM may be used to model other aspects of the processing of quantifiers beside sentence-picture verification. A short theoretical section is devoted to this question. Two potential linking hypotheses between the IPM and processing difficulty during reading are discussed there.

7.1 DIFFICULTY DURING READING

In line with current semantic theory, the IPM assumes that the internal composition of DE comparative quantifiers is more complex than that of UE ones. This may lead one to expect that DE comparative quantifiers cause difficulty during reading in comparison to their UE counterparts. Closely related to this issue, a number of recent studies used event related potentials (ERPs) to test whether the quantifiers *few* and *many* as well as other pairs of UE vs. DE quantifiers are interpreted incrementally (Urbach & Kutas, 2010; Urbach et al., 2015; Freunberger & Nieuwland, 2016; Nieuwland, 2016), as evidenced by the N400 component (see also Fischler, Bloom, Childers, Roucos, &

Perry, 1983). For instance, Nieuwland (2016) presented stimulus sentences like the following.

(7:1) a. Many gardeners plant their flowers during the spring...
b. Many gardeners plant their flowers during the winter...
c. Few gardeners plant their flowers during the spring...
d. Few gardeners plant their flowers during the winter...

In the majority of studies, an incongruency between online N400-effects and offline judgments like plausibility ratings or truth-value judgments was found. While offline ratings always showed a "full crossover interaction" (Urbach & Kutas, 2010), i.e. increased processing cost for 7:1-b vs. 7:1-a, and for 7:1-c vs. 7:1-d, the N400-effects on the critical words, e.g. *spring* vs. *winter*, were reversed only under certain conditions. More specifically, conditions with UE quantifiers like 7:1-a/b always showed congruent online and offline effects whereas in conditions with DE quantifiers this was dependent on several additional factors like, e.g., supporting linguistic context, high cloze probability of the critical words or auditory presentation of the stimulus sentences. These results have been taken to show that DE quantifiers, like *few*, are often interpreted delayed in comparison to their UE counterparts, like *many*.

With regard to the behavioral literature, the question whether DE quantifiers cause disruption during online comprehension that is observable in reading behavior, as stated in the following hypothesis, is still an open question.

Hypothesis 7:2. *As compared to their UE counterparts, DE comparative quantifiers lead to disruption during online comprehension that is due to their semantic processing difficulty and observable in reading behavior.*

Self-paced reading experiments that compared UE vs. DE comparative quantifiers obtained mixed results. On the one hand, Geurts et al. (2010) and Szymanik and Zajenkowski (2013), who used self-paced reading of complete sentences, did not find any significant difference in reading times between sentences that either contained an UE or a DE comparative quantifier. In these studies, the variance was relatively large because the dependent variable were reading times of complete sentences. This may have led to low statistical power. Moreover, on the conceptual side, the reading time data of these experiments are difficult to interpret because of lexical differences between conditions.

On the other hand, Bott et al. (2013) did find increased reading times for DE vs. UE quantifiers. They tested doubly quantified sentences in a word-by-word self-paced reading experiment with moving window presentation (Just, Carpenter, & Woolley, 1982). The effects were found on the final word of the test sentences which also completed the quantifiers' scope. More evidence that points in the same direction was obtained by Bott et al. (n.d.), who compared intransitive sentences with *more than five* vs. *fewer than five* (cf. example 6:18). The latter quantifier led to a slowdown on the sentence-final region which contained the scope of the quantifier. In that study, sentences were, however, only segmented into two regions and, therefore, spillover effects from the regions containing the quantifiers cannot be excluded with certainty. Such spillover effects could simply be due to differences in lexical frequency, for example. In sum, both behavioral and ERP studies offer mixed results concerning the questions whether DE quantifiers lead to on-line processing difficulties in comparison to their UE counterparts.

In order to contribute to this debate, two sets of eye tracking data that compared the online comprehension of sentences with UE vs. DE comparative quantifiers are presented in what follows. Both of these data sets were collected together with the data presented in Experiment 3a. In order to exclude spillover effects a simple "recipe" was used: (1) A buffer region was included between the region containing the quantifier and the region that contained its scope – an AdjP in both cases. (2) Effects on the adjectival region were only interpreted if no differences were observed in the buffer region. In both data sets, evidence was found that the DE quantifiers cause disruption during reading. Towards the end of this chapter, two possible explanations of the observed disruptions in terms of the IPM are discussed briefly.

7.2 EXPERIMENT 3B

As was described above in section 5.4, eye movements were recorded in addition to the verification data reported in Experiment 3a. These are the subject of the present section.

7.2.1 *Methods*

7.2.1.1 *Materials*

Materials were the same as in Experiment 3a. The factor *modifier* (two levels: *more* and *fewer*) was crossed with the factor *numeral* (four levels: *four*, *six*, *eight*, *ten*) which led to eight conditions (see example 5:23). Sentences were segmented into regions of interest (ROIs) (of course, invisible to the participants) as is shown in the following example:

(7:3) a. |$_1$Mehr als vier |$_2$Punkte |$_3$auf dem Bild |$_4$sind blau.
|$_1$More than four |$_2$dots |$_3$on the picture |$_4$are blue.

b. |$_1$Weniger als vier |$_2$Punkte |$_3$auf dem Bild |$_4$sind blau.
|$_1$Fewer than four |$_2$dots |$_3$on the picture |$_4$are blue.

7.2.1.2 *Procedure*

Same as in Experiment 3a.

7.2.1.3 *Participants*

Same as in Experiment 3a.

7.2.1.4 *Predictions*

The present experiment tested Hypothesis 7:2. As was hinted at above, in connection to the study of Bott et al. (n.d.), there is a technical complication in testing this hypothesis. The quantifiers themselves, e.g. *fewer than n* vs. *more than n*, differ in a number of superficial lexical features, such as word length and lexical frequency. Therefore, potential effects on the regions that contain the quantifiers cannot be interpreted safely as reflecting semantic processing difficulty. Moreover, such potential effects may cause spillover effects at the subsequent region. To circumvent this difficulty, the present experiment made use of the recipe mentioned above: Effects on the quantifiers themselves and on following spillover regions are not interpreted. Instead, our analyses exclusively focus on a later sentence region that contained the scope of the quantifiers. Effects on that region are only interpreted in the absence of effects on the regions that directly preceded it.

This prediction that the semantic complexity of the quantifiers surfaces when reading their scope can be justified theoretically: When reading the quantifiers' scope, it is integrated into the semantic representation of the sentence. This presumably involves retrieving the quantifier from memory. Moreover, this is the first position in the sentence at which the quantifier can be fully interpreted, i.e. has received all its arguments (c.f. Bott et al., 2013, n.d.; Nieuwland, 2016).

7.2.1.5 *Statistical analysis*

Before statistical analysis, the eye movement recorded was preprocessed. All fixations below 80*ms* were merged into the nearest neighbor fixation if it was within a radius of 1°. After that, all fixations below 80*ms* or above 1200*ms* (4% of all fixations) as well as fixations that fell outside of any ROI (8.8% of all fixations) were deleted. Next, after visual inspection of the eye movement record, trials with total reading times below 560*ms* were excluded from further analysis. In addition, trials with total reading times of more than 10*s* were also excluded. Together, the last two steps affected 1.29% of all trials. These trials were regarded as contaminated.

The following eye tracking measures were calculated and analyzed (cf. Clifton, Staub, & Rayner, 2004):

- *First pass durations*: total duration of all fixations in a ROI from its first fixation until it is first left, provided that the region was not skipped
- *First pass regression ratios*: proportion of regressive fixations following fixation in a ROI, provided that (*i*) the region was not skipped and that (*ii*) the ROI has not already been fixated and exited
- *Total duration*: total duration of all fixations in a ROI.

First pass durations and total durations were analyzed using linear mixed effects models. First, saturated models were fitted. These included only the factor *modifier* as fixed effect. Each of the noun adjective combinations was coded as an item (cf. section 5.4.1.2). As random effects, the models included random intercepts of participants, items and also of the factor *numeral*. In addition, by-participant, by-item and by-number random slopes were included, i.e. the maximal random effects structure was used but without random correlations

(as suggested by Barr et al., 2013). Next, model comparisons on the basis of the LRT were used to test for significant effects. The proportions of first pass regressions were analyzed with logit mixed effects models (see Jäger, 2008 for discussion of its advantages) using the same procedure. As justified above, analyses were carried out on the sentence-final region and the preceding buffer regions. Predictions were, however, restricted to the final region and the analyses on the buffer regions only served the purpose of excluding spillover effects. All models were fitted using the `lme4` package (Bates, Mächler, et al., 2015) of R (R Core Team, 2016).

7.2.2 *Results*

On the first ROI, *fewer than n* led to longer reading times than *more than n* (mean first pass durations: 623 ms vs. 552 ms; mean total durations: 906 ms vs. 745 ms). This can be explained by superficial differences of the lexical material in that region (e.g. word length). Mean first pass durations, first pass regression ratios and total durations in ROIs #2 through #4 are shown in Figure XL. The results of the statistical analysis are given in Table XLI. As demanded by the above mentioned recipe, there was no effect of the *modifier* on the buffer ROIs, #2 and #3. Therefore, we are allowed to interpret effects on the final ROI. Although on the final ROI first pass durations were numerically longer and first pass regression ratios slightly higher for the DE than for the UE conditions (305 ms vs. 294 ms and 44.3 % vs. 41.9 %, respectively), this did not result in significant effects. Total durations on the final ROI were, however, significantly longer for DE than for UE conditions, albeit the effect was nummerically rather small (346 ms vs. 320 ms).

7.2.3 *Discussion*

On the sentence final region, which contained the color adjective, DE modified numerals of the form *fewer than n* led to significantly longer total durations than UE modified numerals of the form *more than n*. In line with our recipe, no difference was found in earlier regions. Therefore, we interpret the effect to reflect semantic processing difficulty due to the integration of the scope of the quantifiers (i.e. the adjective) into the semantic representation. Although statistically robust, the effect was numerically small. The next section reports an experi-

ment in which UE and DE quantifiers were combined with negation in order to boost the effect (cf. the introduction to the next section which refers to Sherman, 1976 and Vázquez, 1981).

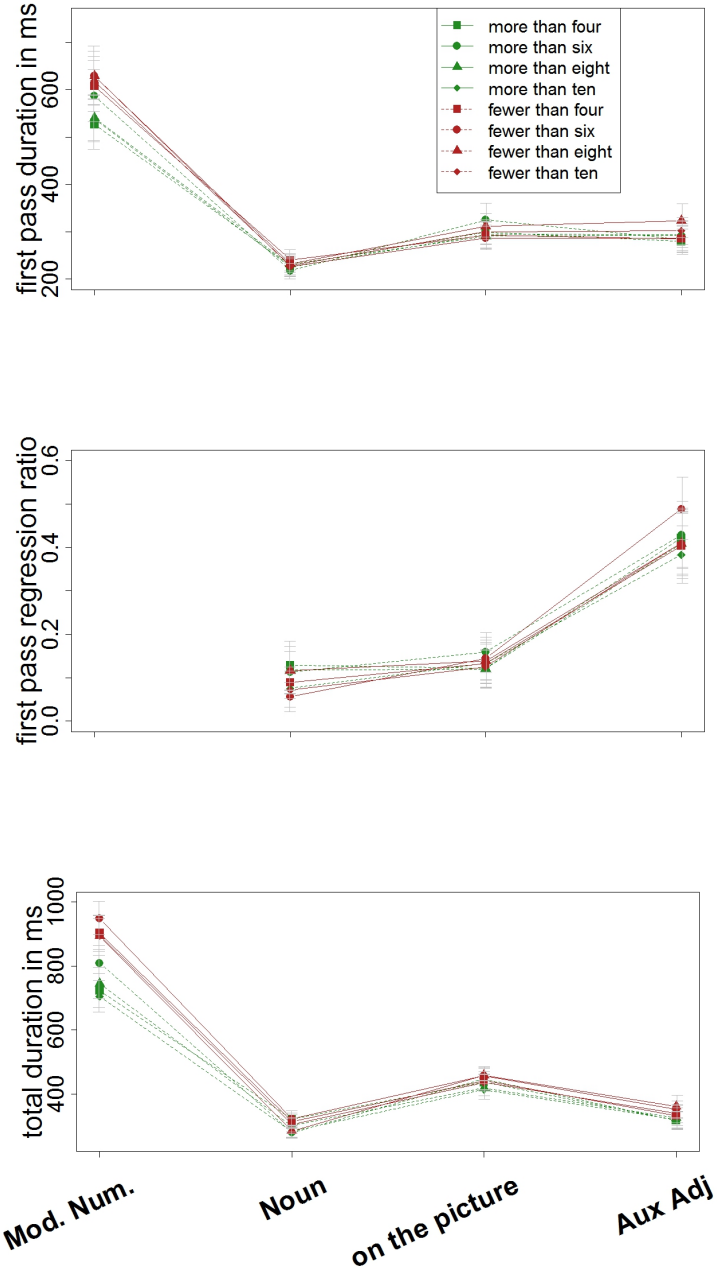


Figure XL. Eye movement measures from Experiment 3b. The depicted means and 95% confidence intervals were calculated from the by-participant means. The legend applies to all four panels.

Table XLI

Statistical analysis of Experiment 3b.

	ROI #2	ROI #3	ROI #4
1st pass duration			
mod	$\chi^2(1) = 1.32, p = .25$	$\chi^2(1) = 0.001, p < .968$	$\chi^2(1) = 1.31, p = .252$
1st pass regression ratio			
mod	$\chi^2(1) = 1.22, p = .269$	$\chi^2(1) = 0.256, p = .612$	$\chi^2(1) = 2.04, p = .152$
total duration			
mod	$\chi^2(1) = 0.553, p = .457$	$\chi^2(1) = 3.01, p = .0830^{\cdot}$	$\chi^2(1) = 5.55, p = .0184^{*}$

Note. mod: modifier; * : $p < .05$;* : $p < .1$.

7.3 EXPERIMENT 5

In the present experiment, eye movements were recorded while participants read scope disambiguated sentences containing either an UE or a DE quantifier in subject position and a possibly negated adjectival predicate in the scope of the quantifier. As in the previous experiment, a truth-value judgment had to be given afterwards, at the end of each trial. We expected that combining the quantifiers with negation would magnify the effect of the direction of entailment. Especially under semantic proposals which assume that the antonym operators FEW and LITTLE contain propositional negation (e.g. I. Heim, 2006; Büring, 2007b), it is expected that the effect of the direction of entailment is magnified in combination with negation. This expectation is based on previous empirical findings which show that multiple negations in a sentence have over-additive effects on processing difficulty (Sherman, 1976; Vázquez, 1981).

The prediction is, however, not limited to proposals that assume a propositional negation operator to be part of the lexical semantics of DE quantifiers. The relation between the direction of entailment and negation is a more general one. Firstly, as was already mentioned above in section 5.2.3, any DE quantifier is the negation of some suitable UE counterpart. Secondly, even if FEW and LITTLE do not contain propositional negation but map degrees to their negative image, as in the IPM, they share some properties with negation: In both accounts antonym operators and overt sentence negation are modifiers in the sense that they do not change the semantic type or syntactic category of their arguments. Moreover, both have the potential to introduce subtle ambiguities (cf. section 5.2.3.2). For these reasons, it can generally be expected that overt negation taxes processing resources that are also needed for the online comprehension of DE comparative quantifiers. Conversely, if the two factors affected independent and serial processing stages, we would expect purely additive effects according to the *additive factors* logic (Sternberg, 1969).

In addition to the direction of entailment, the specific quantifier type was manipulated: Comparative proportional quantifiers (*more than half* and *less than half*) were compared with Aristotelian quantifiers (*every* and *no*). For Aristotelian quantifiers, a decompositional analysis which makes use of covert negation is commonly assumed (see e.g. Jacobs, 1982; Penka & Stechow, 2001; Penka, 2011 but also Abels & Martí, 2010, Geurts, 1996 and de Swart, 2000 for opposing

views). However, with regard to proportional quantifiers such an analysis has only little empirical support. By comparing these quantifier types, the present study thus contributes to a current theoretical debate about the compositional structure of DE as compared to UE quantifiers from a processing perspective.

7.3.1 *Methods*

The trials of the present experiment served as filler trials in Experiments 3a/b.

7.3.1.1 *Materials*

Thirty-two sets of sentences like the example in 7:4 were constructed. The sentences contained a quantifier in subject position and a possibly negated adjectival predicate. The factors *type of quantifier* (levels: *Aristotelian* and *proportional*), *negation* (levels: *present* and *absent*) and *direction of entailment* (levels: *UE* and *DE*) were crossed yielding the eight sentence conditions exemplified in 7:4-a−7:4-h. The Aristotelian quantifiers were *jed-* ('every', UE) and *kein-* ('no', DE). As proportional quantifiers *Mehr als die Hälfte d-* ('more than half of the', UE) and *Weniger als die Hälfte d-* ('less than half of the', DE) were used. Sentences were constructed according to the pattern: *for* Q N *it holds that* PRO *(*NEG*)* AUX ADJ. The clause-boundary was placed between the quantifier and the negation in order to ensure a surface scope interpretation of the sentences (see, e.g., Büring, 1997 for a discussion of scope ambiguities in German sentence with quantifiers and negation and Ruys & Winter, 2011 for a recent discussion of restrictions on scope ambiguities). This way effects of the processing of scope ambiguity could be excluded. The nouns denoted geometrical shapes (*Punkt(e)* ('dot(s)'), *Dreieck(e)* ('triangle(s)'), *Quadrat(e)* ('square(s)') or *Kreuz(e)* ('cross(es)')). The adjectives were *grün* ('green'), *rot* ('red'), *blau* ('blue') or *orange* ('orange'). Thus, there were 16 distinct combinations of nouns and adjectives in total. The vertical lines in the example sentences indicate the ROIs.

(7:4) a. |Auf |jedes Quadrat |trifft zu, |dass es |blau ist.
|on |every square |applies |that it |blue is
'It holds for every square that it is blue.'

b. |Auf |jedes Quadrat |trifft zu, |dass es |nicht blau ist.
|on |every square |applies |that it |not blue is
'It holds for every square that it is not blue.'

c. |Auf |kein Quadrat |trifft zu, |dass es |blau ist.
|on |no square |applies |that it |blue is
'It holds for no square that it is blue.'

d. |Auf |kein Quadrat |trifft zu, |dass es |nicht blau ist.
|on |no square |applies |that it |not blue is
'It holds for no square that it is not blue.'

e. |Auf |mehr als die Hälfte der Quadrate |trifft zu, |dass
|on |more than the half of squares |applies |that
sie |blau sind.
they |blue are
'It holds for more than half of the squares that they are blue.'

f. |Auf |mehr als die Hälfte der Quadrate |trifft zu, |dass
|on |more than the half of squares |applies |that
sie |nicht blau sind.
they |not blue are
'It holds for more than half of the squares that they are not blue.'

g. |Auf |weniger als die Hälfte der Quadrate |trifft zu, |dass
|on |less than the half of squares |applies |that
sie |blau sind.
they |blue are
'It holds for less than half of the squares that they are blue.'

h. |Auf |weniger als die Hälfte der Quadrate |trifft zu, |dass
|on |less than the half of squares |applies |that
sie |nicht blau sind.
they |not blue are
'It holds for less than half of the squares that they are not blue.'

Each sentence was paired with a picture showing objects of the mentioned shape. A target item consisted of a set of eight sentences-picture pairs. In half of these items, the sentence described the picture truthfully. In the other half, it did not. Example pictures are shown in Figure XLII. For each sentence, a new picture with random object positions was generated. Thus, no picture was used twice.

The target items were distributed over the eight lists of Experiment 3a/b using a Latin square design. Each list included 391 additional sentence-picture pairs, yielding a total number of 423 trials. These distractor sentences also contained quantifiers, nouns that described geometrical shapes and color adjectives. They were paired with similar pictures as in the target items. In approximately half of the trials in each list, the sentences matched the pictures and in the other half they did not.

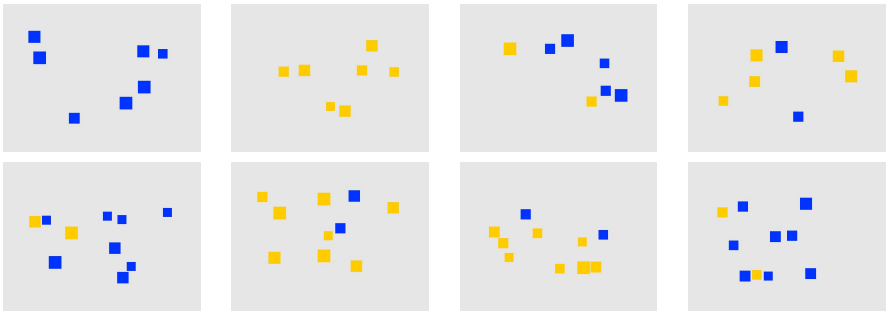


Figure XLII. Example pictures used in Experiment 5. The first row shows pictures that satisfied the sentences' truth conditions. The second row shows pictures that don't. From left to right, pictures of these types were paired with the sentence conditions exemplified in 7:4-a & 7:4-d, 7:4-b & 7:4-c, 7:4-e & 7:4-h and 7:4-g & 7:4-g, respectively.

7.3.1.2 *Apparatus and Procedure*

Were the same as in Experiment 3a/b.

7.3.1.3 *Participants*

Were the same as in Experiment 3a/b.

7.3.1.4 *Predictions*

As explained in the introduction to the present section, it was expected that effects of the direction of entailment are magnified in combination with negation. Under semantic theories which propose a parallel treatment of Aristotelian and comparative quantifiers, for example with regard to their internal composition (e.g. Abels & Martí, 2010), it is predicted that the two types of quantifiers are affected in the same way when combined with negation. Therefore, the mentioned type of uniform approach would be in need for an explanation if over-additive effects of the *direction of entailment* and *negation* were found within the Aristotelian but not within the proportional quantifiers.

In addition to these predictions, a number of trivial lexical effects were expected. For example, the second ROI, which contained the quantifiers, was substantially longer in conditions with proportional than with Aristotelian quantifiers. Moreover, UE and DE quantifiers also differed in length and presumably also in frequency of occurrence. These factors led us to expect differences in eye tracking

measures between conditions in the quantifier ROI. Furthermore, effects due to the presence or absence of negation were expected in the sentence-final ROI.

What is more, we cannot exclude the possibility that lexical effects carry over to the 'buffer' regions (ROI #3 and ROI #4) due to spillover or preview. For this reason the recipe from the previous experiment was used again and a prerequisite to interpreting effects involving the *direction of entailment* in the sentence final ROI was formulated: Such effects would only be interpreted as reflecting semantic processing if they were absent on at least one of the two buffer regions. This way, effects of semantic processing difficulty can be distinguished from superficial lexical ones.

7.3.1.5 *Statistical Analysis*

Preprocessing of the eye movement record proceeded exactly as described in Experiment 3b. Total durations were not analyzed in the present experiment. In addition to the eye-tracking measures mentioned above in section 7.2.1.5, the following eye-tracking measures were analyzed in the present experiment (cf. Clifton et al., 2004):

- *Regression path durations*: total duration of all fixations from the first fixation in a ROI until it is exited in a progressive manner, provided that the region was not skipped
- *Second pass duration*: total duration of all fixations in a ROI following the initial first pass time, including zero durations if a region is not re-fixated

First pass durations, regression path durations and second pass durations were analyzed using linear mixed effects models. First, saturated models were fitted. As fixed effects these included the factors *type of quantifier*, *direction of entailment*, *negation* and their interactions. As random effects they included random intercepts of participants and items as well as by-participant and by-item random slopes of all the fixed effects. As above, model comparisons were used to test for significant effects. The proportions of first pass regressions and the truth-value judgments were analyzed with logit mixed effects models. In the analysis of eye movements, each of the noun adjective combinations was coded as an item. In the analysis of the judgment data, truth values were also taken into account in the coding of items. The procedure outlined by Levy (2014) was used to test for significant

main effects in the presence of higher order interactions. In case of non-convergence of the statistical models, the random effects structure was simplified (cf. Barr et al., 2013). This happened on two occasions. In the analysis of first pass durations in ROI #5 and of the truth value judgments the by-item random slopes of the three-way interaction was dropped.

Because of our recipe for the interpretation of potential effects, first pass durations, first pass regression ratios, regression path durations and second pass durations were analyzed in ROIs #2 through #5. In the following results section, all effects are mentioned that are significant at $\alpha = .05$. Predicted effects are highlighted in the text. Moreover, it is also highlighted if effects can be explained based on superficial lexical features of the stimulus material. In the discussion below, the focus is on effects without an obvious lexical explanation and in particular on those mentioned in the predictions.[1]

7.3.2 *Results*

7.3.2.1 *Eye movements*

Condition means and 95% confidence intervals of the eye tracking measures in all the ROIs are shown in Figure XLIII. Results of the statistical analysis are shown in Tables XLV and XLV. As expected, there were a number of trivial effects. For example, proportional quantifiers led to longer reading times on the second ROI than Aristotelian ones simply because this region was longer and contained more words in the former conditions. Unsurprisingly, this effect was highly significant across reading time measures. Moreover, lexical differences between conditions may also account for an interaction of the *direction of entailment* and the *type of quantifier* in the second region and the effect of the *type of quantifier* in the first pass regression ratios in the following ROI. Similarly, a number of significant effects were found on the fourth and on the fifth ROI that involve the factor

[1] Following current practice in psycholinguistic studies of eye movements during reading, no α-correction was performed. Lately, von der Malsburg and Angele (2015) did, however, recommend to do so. The predictions for the present experiment concerned eye movements that took place after the first visit of the sentence-final ROI. Relevant measures are first pass durations, first pass regression ratios and regression path durations on the final ROI as well as second pass durations on all ROIs (most importantly the second and final regions). In total this adds up to a maximum of seven relevant measures with mutual stochastic dependencies. On the basis of the provided information, the reader may form his own judgment as to whether effects should be considered significant.

negation. Some of these effects are most likely due to the one additional word in the negated conditions that was fixated in ROI#5 and lay within parafoveal preview of fixations in ROI#4. In addition, the adjective lay within the *word identification span* (Rayner, 1998) when ROI#4 was fixated in the non-negated conditions whereas it was beyond the word identification span in the negated conditions. In the former case, regressions that are typically launched from the end of a sentence may have been launched earlier because the end of the sentence was already predictable. Finally, in the sentence final ROI, the singular vs. plural form of the auxiliary in the proportional vs. Arsitotelian conditions, respectively, showed clear effects in the eye movement record.

In addition to these expected but trivial effects, our central predictions were borne out. In particular, let us focus on the first pass regression ratios and regression path durations in the sentence final ROI. In both measures, a significant interaction of the *direction of entailment* and *negation* as well as a main effect of the *direction of entailment* were observed. The three-way interaction was, however, far from significant in both measures. A detailed picture of these eye tracking measures is shown in panels a and b of Figure XLIV, respectively. As predicted, DE quantifiers led to a higher ratio of first pass regressions and longer regression path durations on average than UE quantifiers. Moreover, these increases were larger in negated than in non-negated conditions. To resolve the interaction, conditions with and without negation were analyzed separately. Within the non-negated conditions, the main effect of the *direction of entailment* was neither significant in the first pass regression ratios ($\chi^2(1) = 0.209, p = .648$) nor in the regression path durations ($\chi^2(1) = 2.62, p = .105$). In contrast, in the negated conditions, the DE quantifiers led to reliably higher ratios of regressions ($\chi^2(1) = 8.69, p = .0032$) and reliably longer regression path durations ($\chi^2(1) = 18.8, p < .001$) than DE quantifiers.

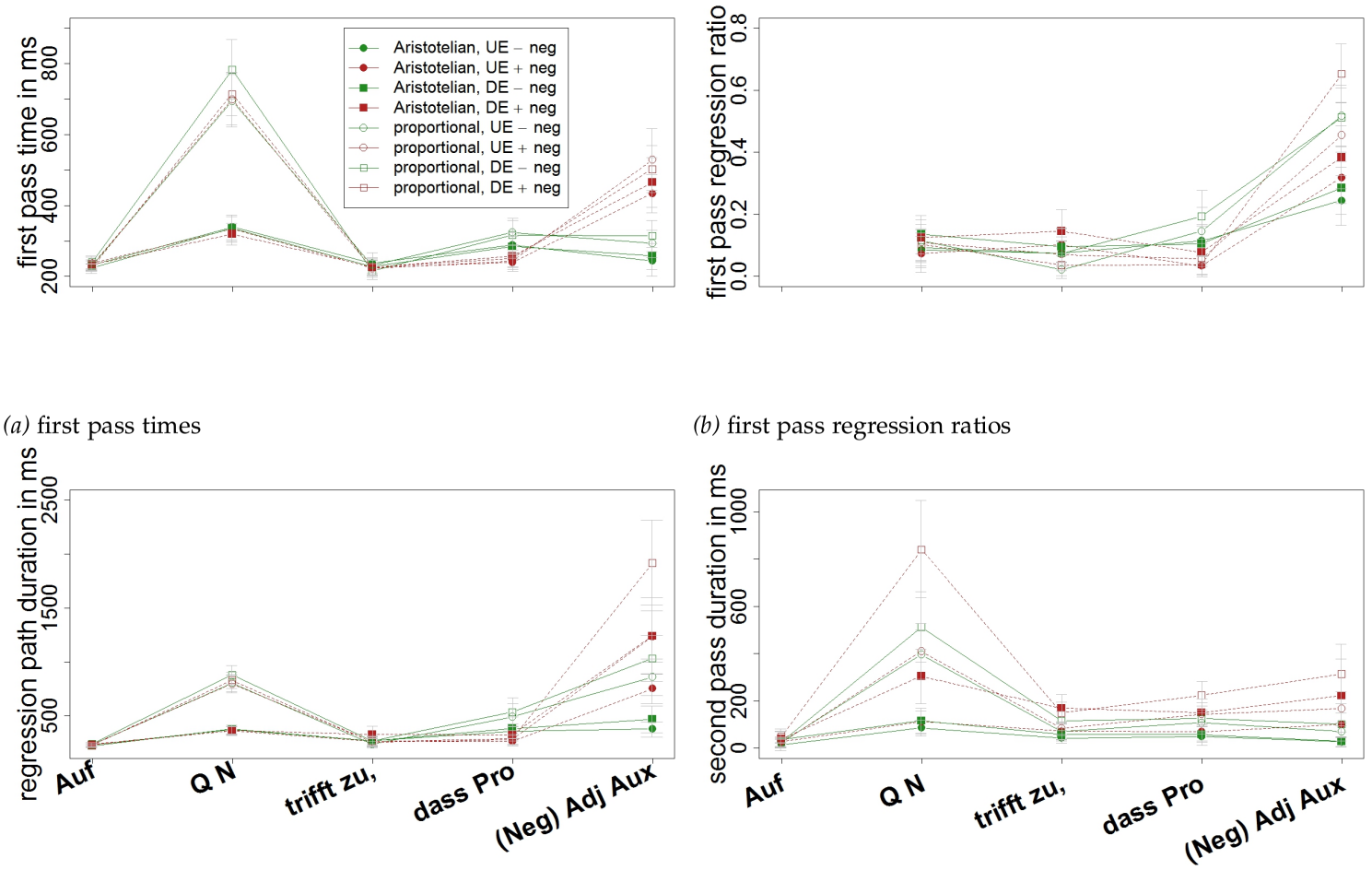
These effects carried over to the second pass durations. Specifically, let us have a look at the second pass durations in regions #2 and #5. These are shown in panels c and d of Figure XLIV. In both of these regions, there was again a significant interaction between the *direction of entailment* and *negation*. Separate analyses of negated and non-negated conditions, revealed that DE quantifiers led to significantly longer second pass times in the negated conditions (ROI#2: $\chi^2(1) = 19.9, p < .001$; ROI#5: $\chi^2(1) = 6.05, p = .0139$) but not in non-negated conditions (ROI#2: $\chi^2(1) = 4.53, p = .103$; ROI#5:

$\chi^2(1) = 0.317, p = .573$). Moreover, the same pattern of effects was observed in the fourth region.

It is important to note here that, in contrast to the effects discussed at the beginning of this section, none of the effects mentioned in the previous two paragraphs has an obvious explanation in terms of superficial lexical differences. Moreover, there were no effects involving the *direction of entailment* during first pass reading in the buffer regions, #3 and #4. Thus, according to our recipe we are allowed to interpret these effects.

7.3.2.2 *Judgments*

The proportions of errors in the different conditions and approximate 95% confidence intervals are shown in Figure XLVII. Overall, negated sentences were judged erroneously more often than non-negated ones (14% vs. 4%) producing a significant main effect (*negation*: $\chi^2(1) = 33.5, p < .001$). Furthermore, proportional quantifiers led to significantly more errors than Aristotelian ones (14.5% vs. 4.2%). This also led to a reliable effect (*type of quantifier*: $\chi^2(1) = 23.7, p < .001$). Finally, DE quantifiers led to more errors than UE ones (14.3% vs. 4.4%). This was reflected in a reliable main effect of *the direction of entailment* ($\chi^2(1) = 24.6, p < .001$). There were no significant interactions in the proportions of errors. In particular, the interaction between the *direction of entailment* and *negation* was far from significant ($\chi^2(1) = 0.586, p = .444$).



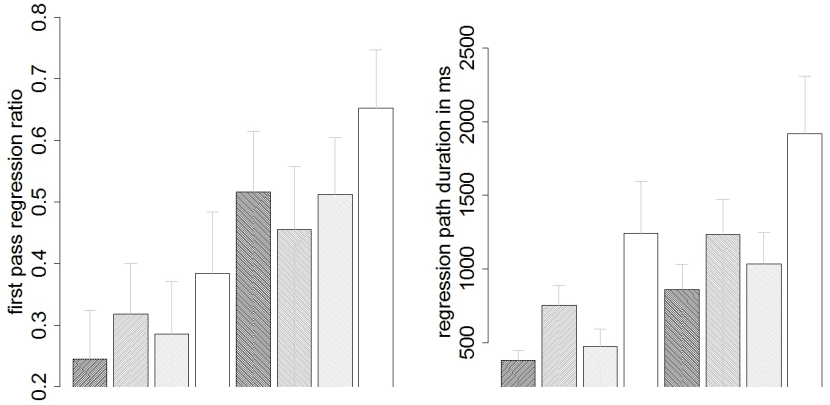
(a) first pass times

(b) first pass regression ratios

(c) regression path durations

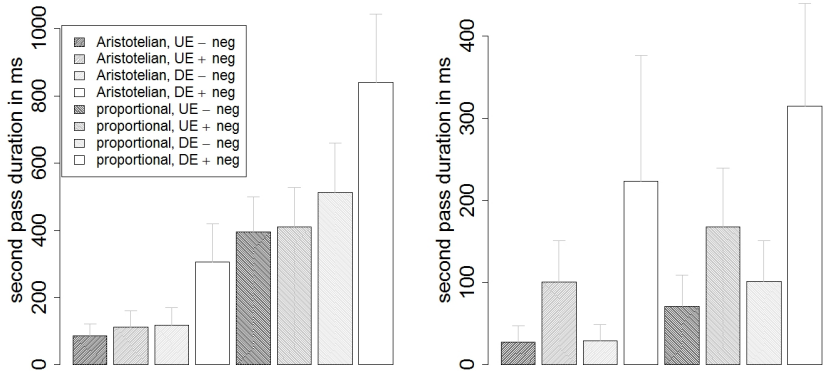
(d) second pass durations

Figure XLIII. Eye tracking measures obtained in Experiment 5. The depicted means and 95% confidence intervals were calculated from the by-participant means. The legend in panel a applies to all four panels.



(a) first pass regression ratios, ROI #5

(b) regression path durations, ROI #5



(c) second pass durations, ROI #2

(d) second pass durations, ROI #5

Figure XLIV. Selected eye tracking measures of Experiment 5. Means and 95% confidence intervals based on participant means are depicted. The legend in panel b applies to all four panels.

Table XLV

Statistical analysis of 1st pass durations and 1st pass regression ratios from Experiment 5.

	ROI #2	ROI #3	ROI #4	ROI #5
1st pass durations				
doe	–	–	–	–
neg	–	–	$\chi^2(1) = 18.6, p < .001^{***}$	$\chi^2(1) = 39.9, p < .001^{***}$
toq	$\chi^2(1) = 102, p < .001^{***}$	–	–	$\chi^2(1) = 10.2, p = .00141^{**}$
doe×*neg*	–	–	–	–
doe×*toq*	$\chi^2(1) = 4.87, p = .0274^{*}$	–	–	–
neg×*toq*	–	–	–	–
doe×*neg*×*toq*	–	–	–	–
1st pass regression ratios				
doe	–	–	–	$\chi^2(1) = 4.56, p = .0328^{*}$
neg	–	–	$\chi^2(1) = 11.6, p < .001^{***}$	$\chi^2(1) = 6.88, p = .00873^{***}$
toq	–	$\chi^2(1) = 12.9, p < .001^{***}$	–	$\chi^2(1) = 32.4, p < .001^{***}$
doe×*neg*	–	–	–	$\chi^2(1) = 6.11, p = .0135^{*}$
doe×*toq*	–	–	–	–
neg×*toq*	–	–	–	–
doe×*neg*×*toq*	–	–	–	–

Note. doe: direction of entailment; *neg*: negation; *toq*: type of quantifier.

Table XLVI

Statistical analysis of regression path durations and 2st pass durations from Experiment 5.

	ROI #2	ROI #3	ROI #4	ROI #5
regression path durations				
doe	–	–	–	$\chi^2(1) = 18.3, p < .001^{***}$
neg	–	–	$\chi^2(1) = 14.5, p < .001^{***}$	$\chi^2(1) = 28.4, p < .001^{***}$
toq	$\chi^2(1) = 107, p < .001^{***}$	–	$\chi^2(1) = 7.14, p = .00756^{**}$	$\chi^2(1) = 27.7, p < .001^{***}$
doe×*neg*	–	–	–	$\chi^2(1) = 7.88, p = .00499^{**}$
doe×*toq*	–	–	–	–
neg×*toq*	–	–	$\chi^2(1) = 9.89, p = .00167^{**}$	–
doe×*neg*×*toq*	–	–	–	–
second pass durations				
doe	$\chi^2(1) = 21.1, p < .001^{***}$	$\chi^2(1) = 17.55, p < .001^{***}$	$\chi^2(1) = 11.9, p < .001^{***}$	$\chi^2(1) = 5.59, p = .0181^{**}$
neg	$\chi^2(1) = 13.1, p < .001^{***}$	$\chi^2(1) = 15.4, p < .001^{***}$	$\chi^2(1) = 11.9, p < .001^{***}$	$\chi^2(1) = 12.3, p < .001^{***}$
toq	$\chi^2(1) = 37.2, p < .001^{***}$	–	$\chi^2(1) = 17.3, p < .001^{***}$	$\chi^2(1) = 9.65, p = .0019^{**}$
doe×*neg*	$\chi^2(1) = 10.0, p = .00157^{**}$	–	$\chi^2(1) = 7.1, p = .00761^{**}$	$\chi^2(1) = 4.86, p = .0276^{*}$
doe×*toq*	$\chi^2(1) = 6.68, p = .00973^{**}$	–	–	–
neg×*toq*	–	–	–	–
doe×*neg*×*toq*	–	–	–	–

Note. doe: direction of entailment; *neg*: negation; *toq*: type of quantifier.

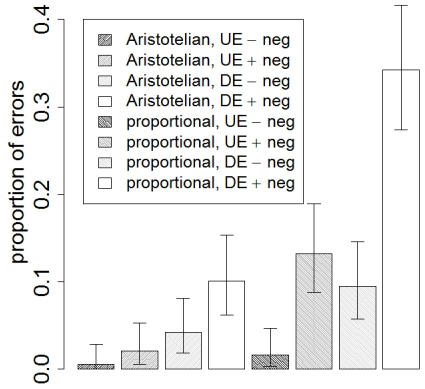


Figure XLVII. Proportions of errors per condition obtained in Experiment 5 and 95% confidence intervals estimated under the assumption of independent samples using the exact method.

7.3.3 *Discussion*

The eye movements were generally as expected: Until the sentence final region was first reached, the observed effects exclusively reflected superficial lexical differences between conditions. Analogously to the previous experiment, the earliest point at which semantic properties – the direction of entailment, in particular – unambiguously affected eye tracking measures was in the first pass regression ratios in the final region. In the negated conditions, DE quantifiers led to substantially higher regression ratios than UE ones. Since no such effects were found in the buffer regions we are entitled – according to our recipe – to interpret the effect as an indication of enhanced semantic processing difficulty.

As expected on the basis of assumptions about the internal composition of DE quantifiers, this increased processing difficulty was strongly pronounced if overt negation was combined with DE quantifiers. It also affected subsequent eye-tracking measures. Regression path durations of the final region and second pass durations in all regions except #1 and #3 were all exceptionally long if DE quantifiers were combined with negation. This confirms the prediction that the direction of entailment has a larger effect in negated than in non-negated sentences. In particular, the results support the hypothesis that the direction of entailment affects processing components that are also involved in the processing of negation. Specifically, decompositional semantic analyses of DE quantifiers led us to expect the observed type of interaction.

In the judgment data, cumulative effects of negation and the direction of entailment were observed. Both, the presence of negation and of downward entailing quantifiers increased the proportions of errors. In addition, proportional quantifiers led to more errors than Aristotelian ones. Thus, all three manipulated factors have an influence on processing difficulty spanning a range from nearly perfect performance to performance near chance level. A tentative conclusion is that in the most difficult conditions comprehension did not succeed and it was in many trials of these conditions impossible to derive a compositional interpretation of the sentences.

7.4 POTENTIAL LINKING HYPOTHESES

The results of the previous two experiments give us reason to believe that, in comparison to their UE counterparts, DE comparative quantifiers lead to enhanced processing difficulty during comprehension (cf. also Bott et al., 2013, n.d. but also the ERP studies of Urbach & Kutas, 2010, Urbach et al., 2015, Freunberger & Nieuwland, 2016 and Nieuwland, 2016). Cases in which DE comparative quantifiers are combined with negation appear to be especially demanding. What may be potential linking hypotheses between the IPM and observations like these? Three speculative possibilities are highlighted here.

Firstly, although it is a model of verification and falsification, the IPM may, in principle, also be used to model how expectations about the state of the world can be derived from the meaning of a sentences. One way to model this kind of process using the IPM is to 'run the verification process in reverse', so to speak (cf. Figure XXIX): In the IPM it was assumed that (1) a representation of the symbolic numeral is retrieved from memory, (2) a representation of numerosity is computed on the basis of visual input, and (3) a truth value judgment is computed by a process that involves comparison of the two. Reversely, we may assume that (1′) a representation of the symbolic numeral is retrieved from memory, (2′) the hearer takes the utterance to be true, i.e. trusts the speaker, and (3′) a representation of numerosity is computed on the basis of the two (cf. Lipinski et al., 2012 for related ideas).

In fact, when we introduced the IPM, it was argued that the computation performed by the comparative morpheme ER is neurobiologically plausible because similar computations have been proposed to implement multisensory integration in the brain (Pouget & Sejnowski, 1997; Dayan & Abbott, 2001). What happens in the latter kind of computation is, somewhat simplified, that one representation is derived or inferred from two others. Which are known and which is inferred does not matter. An example was to infer a representation of an object's position in head-centered coordinates from representations of its eye-centered coordinates and the viewing angle. Adopting this kind of mechanism, it is, in principle, possible to derive a representation that encodes an expectation of how many target objects there are. In a trial of a sentence-picture verification experiment, this would correspond to an expectation about the visual stimulus that is presented (related is also the discussion whether population codes can and do

encode probability distributions see e.g. Zemel, Dayan, & Pouget, 1998; Pouget, Dayan, & Zemel, 2003).

Since DE comparative quantifiers are assumed to involve an additional processing step of scale-reversal, it is conceivable that this kind of process takes longer in the DE than in the UE variant. Moreover, it is not surprising that semantic processing difficulty surfaces when the scope of the quantifier is read because (under canonical word order as used in the above experiments) this is the first position in the sentence where concrete expectations can be derived. Furthermore, the interaction with overt sentence negation can receive a straightforward explanation in this approach. We only need to assume that the process of scale-reversal and some process that is involved in the comprehension of negation tax a common scarce processing resource.

The second possibility is essentially a 'parsing approach.' It is based on the derivation of a symbolic meaning representation. We may think of this in terms of the CCG derivations discussed in sections 5.3.1 and 6.2.1, but we may also think of it in terms of the construction of logical form representations as, for example, shown in appendix A.2. As was discussed in section 5.2.3.2 the antonym operators FEW and LITTLE introduce the potential for subtle ambiguities. However, in the end, sentences like those investigated in the previous experiments turn out completely unambiguous. Now, it is conceivable that the presence of the antonym operator and the way it can interact with other parts of the semantic representation leads the parser to consider parsing decisions that eventually turn out impossible. Furthermore, it is plausible that the earliest point at which the parser can get rid of the impossible alternatives is when the quantifier has received both of its arguments. For example, consider the derivations in section 5.3.1, in which FEW can also combine with MANY instead of ER. In fact, recall that this is the explanation Rullmann (1995) gave for the Seuren-Rullmann ambiguity. If that happens in simple sentences with comparative quantifiers, the result will always be a tautology. Thus, this parse of the sentence is ruled out on semantic grounds, which may only be possible at the end of the sentence or at least after the quantifier has received both of its arguments. That the pruning of parsing alternatives leads to slow down is a common assumption (see e.g. Hale, 2016 for discussion). Under this approach, the observed interaction between the direction of entailment and negation is not surprising because of the combinatorial possibilities. A piece of evidence that supports this type of parsing account was provided

by fMRI study of S. Heim et al. (2012) who found that DE compartive quantifiers lead to increased activity in Broca's area (BA45), which is usually associated with syntactic processing.

The third possibility is the one proposed by Bott et al. (2013, n.d.). In their account increased processing difficulty during online comprehension is attributed to the preparation of a more complex verification algorithm. Concretely, their complex expansion operation (cf. section 6.3.1) consists of more rules that have to be retrieved from long term memory and retained in working memory during the verification task. Moreover, the complex expansion operation depends on what Bott et al. call negative in addition to positive predicate instances. The latter feature renders it possible that DE or rather "empty-set" quantifiers are especially difficult to process when the predicate is negated.

Of course, all of these three possibilities are not worked out, yet. Moreover, they are certainly not the only possibilities to explain the observed effects. Future research has to show which, if any, of these alternative linking hypotheses is on the right track.

8

CONCLUSIONS

The present work discusses several case studies in order to address the question whether and how the semantic processing difficulty of quantified sentences can be modeled based on semantic theory. The general approach is to amend rather uncontroversial aspects of semantic theory with minimal and well-motivated processing assumptions. The main focus is on sentence-picture verification but some connections to comprehension processes are also drawn. Some of the main results are recapitulated in this concluding section.

Inspired by previous work on the computational complexity of polyadic quantification in natural language, chapter 3 investigates reciprocal sentences with quantificational antecedents. It is asked whether the computational complexity of the verification problem induced by different readings of such sentences restricts which interpretations are viable. For certain quantificational antecedents, the verification problem that corresponds to their logically strongest interpretation is computationally intractable. Complementing such complexity analyses chapter 3 presents hypotheses about how the language processor reacts when faced with intractable verification problems. On the basis of these hypotheses predictions about the comprehension and verification of reciprocal sentences are derived. One such hypothesis states, for example, that intractable interpretations are always avoided by a shift in meaning towards a logically weaker alternative because the stronger alternatives simply cannot be expressed in the relevant fragment of the 'language of thought.' Contrary to this hypothesis, empirical data that are presented in chapter 3 show that intractable readings do occur, but in order to verify them against a specific context people rely on certain guessing strategies.

Motivated by these results, a closer look is taken at concrete procedures that are employed to verify or falsify quantificational sentences. Chapter 5 discusses one particularly challenging test case: the verifi-

cation and falsification of UE and DE comparative modified numerals of the form *more than n* and *fewer than n*. As previous experimental work has shown, the DE versions may lead to increased difficulty as compared to their UE counterparts. Since existing processing models have difficulty explaining these effects, several alternative amendments of existing models are discussed that can account for the experimental data. Moreover, an integrated processing model (IPM) is developed that combines semantic theory with insights from cognitive psychology. In line with the interface transparency thesis, this model implements a transparent interface between semantic representations and models of decision processes involved in the comparison of numerosities. Thereby, previous approaches are extended in two respects. Firstly, the compositional fine structure of the studied quantifiers is taken into account, in particular that the DE variants contain an additional semantic operator that has the effect of scale-reversal. Secondly, truth evaluation is described as a stochastic decision process the running time and outcome of which may be affected by various factors. Initial support for the proposed model comes from two experiments that are described in the empirical part of the chapter. In addition, it is discussed what potential relations may exist between the IPM and automata theoretic analyses.

Chapters 6 and 7 discuss generalizations of the IPM to (1) other quantifiers and (2) processes during online language comprehension. With regard to (1), several possible extensions of the model are discussed briefly and one example is worked out in some detail. Moreover, it is discussed how potential extensions may relate to existing experimental findings. One aspect that may be worth highlighting is a discussion of how the IPM relates to so-called empty-set effects during the verification of DE quantifiers, i.e. disproportionate processing difficulty that is observed when no target objects are present. It is shown that the proposed IPM naturally explains such effects. Regarding (2), two eye tracking experiments are reported that demonstrate increased difficulty of DE quantifiers already during reading. It is speculated how the IPM may be extended to model comprehension in addition to verification processes.

A general conclusion that can be drawn from the present work is that theory in semantics and neighboring disciplines has evolved to a point where it is possible to develop integrated processing models. But of course, not all approaches to semantics are equally well-suited to be developed into processing models. And, moreover, not

all auxiliary processing assumptions and linking hypotheses provide us with adequate explanations or accurate predictions of experimental data. It is possible that different theoretical approaches may prove useful for different aspects of semantic processing. In some cases it may even be necessary to modify some of the fundamental assumptions of the underlying semantic theory in order to connect semantic theory to processing data obtained in psycholinguistic experiments. Hopefully, the approach taken in the present work will be useful to model and predict outcomes of future psycholinguistic experiments and, in addition, will also contribute to our understanding of natural language in general.

A

EXAMPLE DERIVATIONS

A.1 THE SEUREN-RULLMANN AMBIGUITY

In section 5.2.3.2, an argument for decomposition of *less* into *-er little* was presented. It was based on the Seuren-Rullmann ambiguity in A:1. Central to the argument was that the ambiguity disappears with *heavier*, which indicates that it is not due to an ambiguity in the *than*-phrase (but cf. Beck, 2012b).

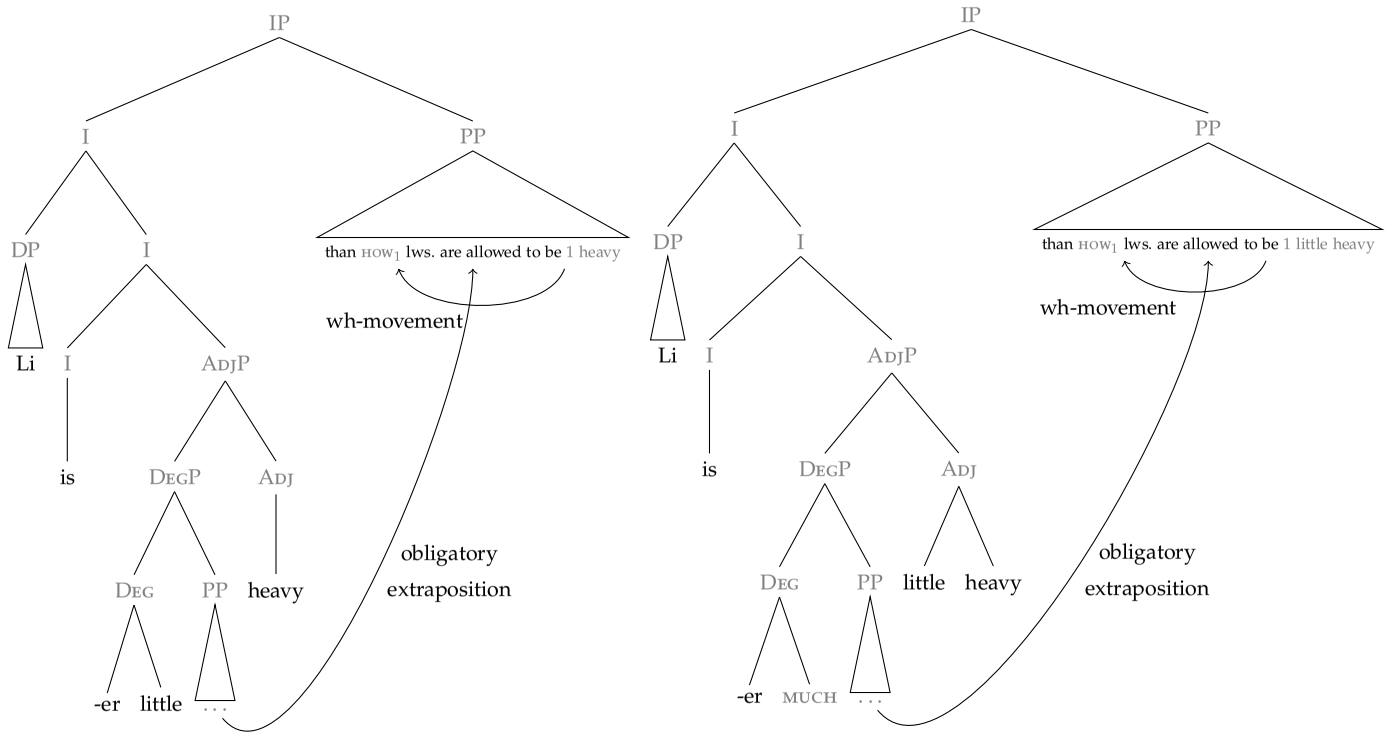
(A:1) Li is less heavy than lightweights are allowed to be.

a. Li's weight is below the maximum permitted weight for lightweights. (maximum reading)

b. Li's weight is below the minimum permitted weight for lightweights. (minimum reading)

In particular, the argument was based on an explanation Rullmann (1995) gave for the ambiguity. Rullmann did not work out the compositional details of his explanation and in section 5.2.3.2 they were omitted as well. The general idea is not repeated here, but the present section describes two alternative compositional derivations of the ambiguity. The first one is that proposed by Büring (2007a, 2007b). The second is a slightly modified version of the former that incorporates some of Rullmann's original assumptions. [1] For simplicity, we use the same syntactic structures in both derivations although there are differences in what structures Rullmann and Büring assumed. The syntactic structure of the maximum reading in shown in Figure XLVIIa and that of the minimum reading in Figure XLVIIIb.

[1] The analysis of I. Heim (2006) is not discussed here because it depends on the assumption that there is a second *little* in the *than*-clause and this assumptions seems difficult to defend when we want to apply the analysis to comprative quantifiers, like e.g. *less than half*, which is the main reason why we discuss ordinary comparatives here.



(a) Maximum reading *(b)* Minimum reading

Figure XLVIII. Possible syntactic structures of the two readings that constitute the Seuren-Rullmann ambiguity.

A.1.1 *Büring's account*

Büring assumes the interpretations below. Additionally, he uses a semantically vacuous MUCH (for discussion see e.g. Wellwood, 2015). Under these assumptions, we derive the maximum reading straightforwardly (Figure XLIXa, cf. section 2.6).

$$\begin{aligned}\llbracket \text{little} \rrbracket_B &= \lambda P.\lambda d'.\neg P\ d' \\ \llbracket \text{heavy} \rrbracket_B &= \lambda d.\lambda x.\mathbf{weight}(x) \geq d \\ \llbracket \text{er} \rrbracket_B &= \lambda \mathcal{T}.\lambda P.\lambda Q.\exists d'(\neg \mathcal{T}(P)\ d' \wedge \mathcal{T}(Q)\ d')\end{aligned}$$

To compose gradable adjectives with *little* Büring uses what he calls the "κ-combinator," which he motivated elsewhere (Büring, 2005). For our purposes, it can be defined as: $\kappa := \lambda R.\lambda \mathcal{S}.\lambda s.\mathcal{S}(\lambda r.R\ r\ s)$. Gradable adjectives then compose with *little* as follows, which allows us to derive the minimum reading (Figure XLIXb).

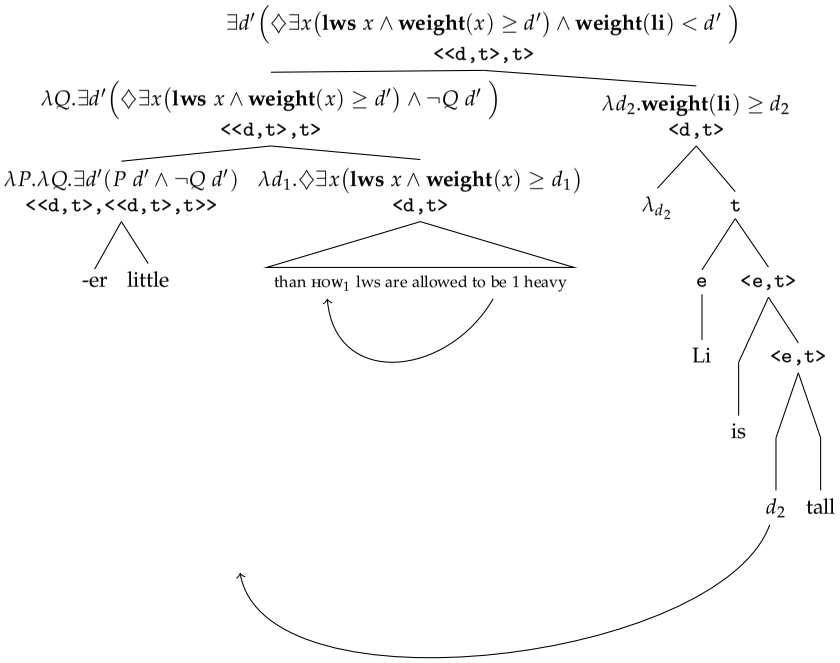
$$\left\llbracket \begin{array}{c} \text{little} \\ \diagup\ \diagdown \\ \kappa \quad \text{heavy} \end{array} \right\rrbracket = \llbracket \text{little} \rrbracket \left(\kappa \left(\llbracket \text{heavy} \rrbracket\right)\right) = \lambda s.\lambda d'.\mathbf{weight}(s) < d'.$$

A.1.2 *Compositional Version of Rullmann's account*

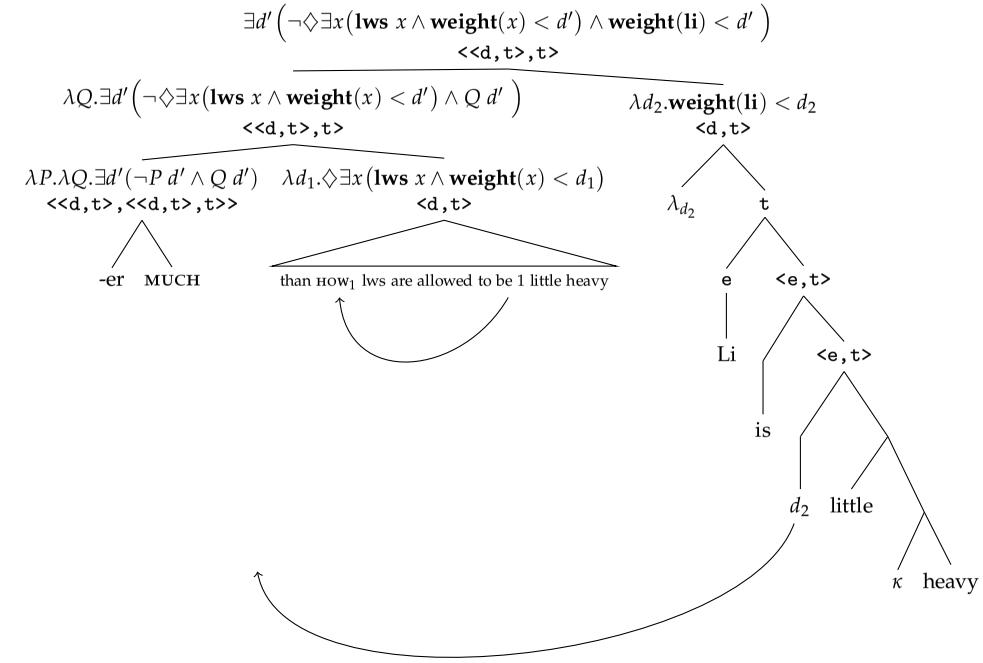
To obtain a compositional version of Rullmann's (1995) account, we make a few changes to the lexicon. The interpretation of *little* is based on I. Heim's (2006) suggestion to implement Rullmann's proposal. That of *-er* is a slightly modified version of what Rullmann suggested.

$$\begin{aligned}\llbracket \text{little} \rrbracket_R &= \lambda P.\lambda d'.P(-d') \\ \llbracket \text{heavy} \rrbracket_R &= \lambda d.\lambda x.\mathbf{weight}(x) = d \\ \llbracket \text{er} \rrbracket_R &= \lambda \mathcal{T}.\lambda P.\lambda Q.\exists d'\Big(\mathcal{T}\big(Q\big)\ d' \wedge \mathcal{T}\big(\lambda d.d' > d\big)\big(\textsc{max}(P)\big)\Big)\end{aligned}$$

This allows us to derive the two readings completely parallel to how it is done in Büring's account. Of course, we have glossed over some difference between the two accounts here. But we have nevertheless seen that Rullmann's core assumptions can be integrated into a derivation of the ambiguity that is as compositional as other alternatives.

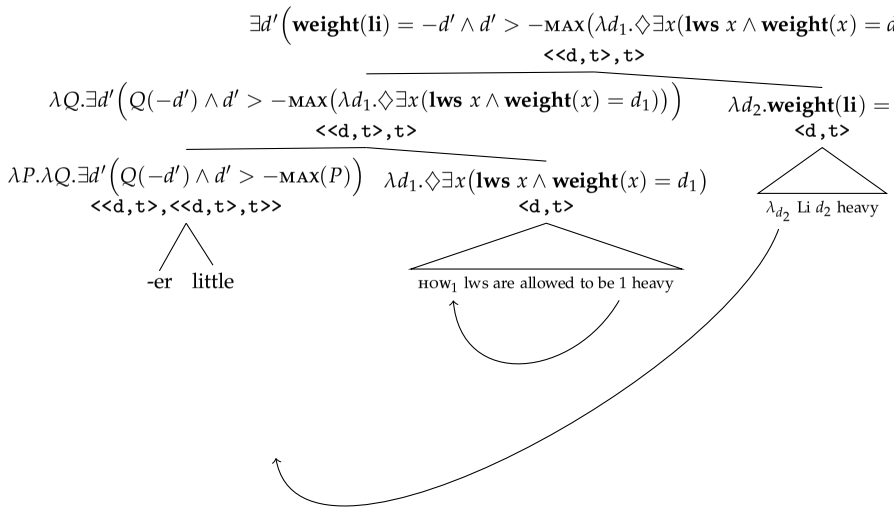


(a) Maximum reading

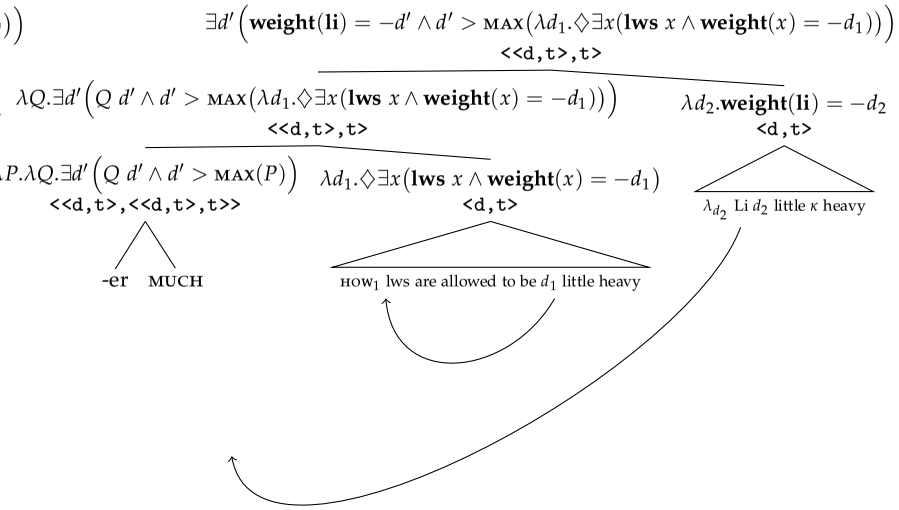


(b) Minimum reading

Figure XLIX. Büring's account



(a) Maximum reading, equivalent to:
$\mathbf{weight}(\mathbf{li}) < \textsc{max}(\lambda d.\Diamond\exists x(\mathbf{lws}\ x \wedge \mathbf{weight}(x) = d))$

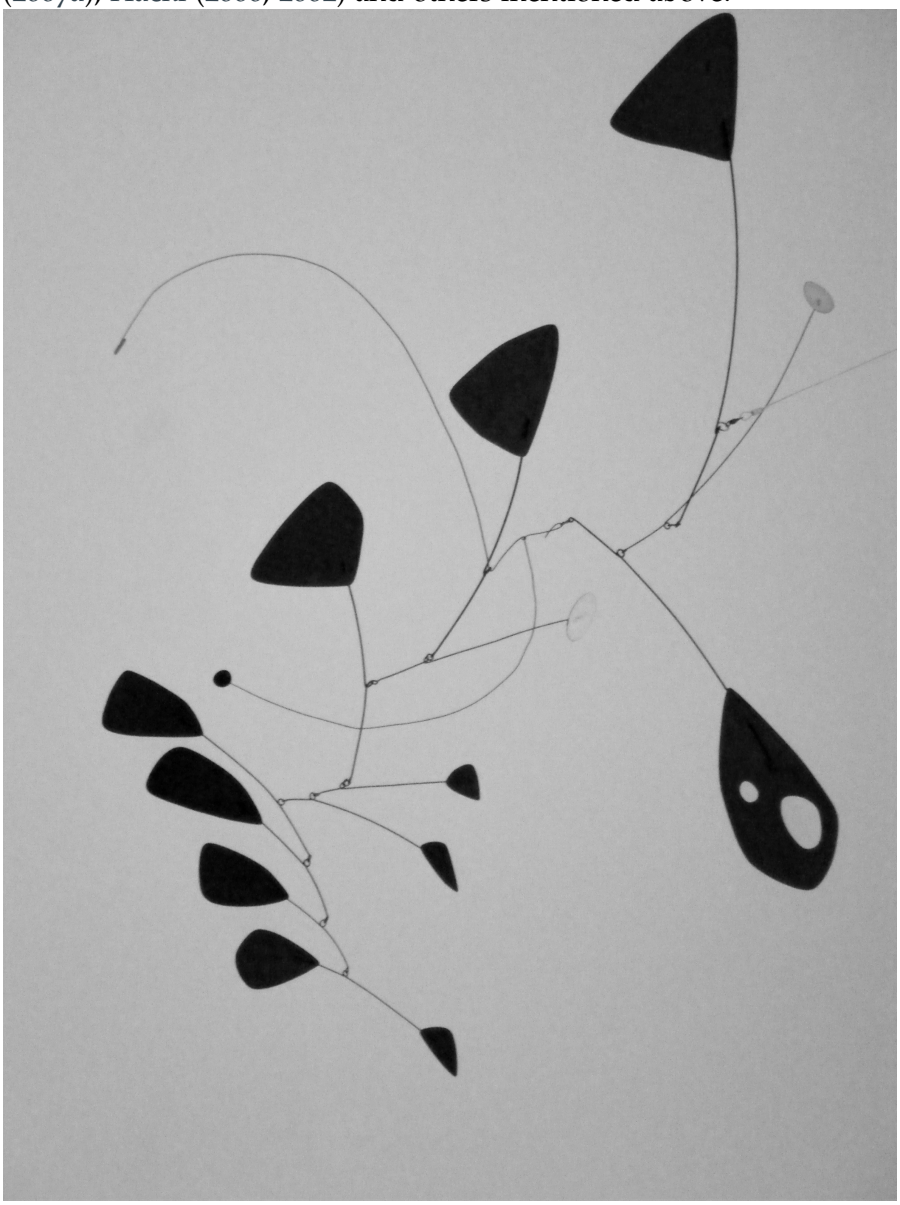


(b) Minimum reading, equivalent to:
$\mathbf{weight}(\mathbf{li}) < \textsc{min}(\lambda d.\Diamond\exists x(\mathbf{lws}\ x \wedge \mathbf{weight}(x) = d))$

Figure L. Compositional version of Rullmann's account

A.2 COMPARATIVE MODIFIED NUMERALS

In section 5.2.3.2, it was briefly discussed how the semantics of ordinary comparatives, which involve gradable adjectives (see section 2.6), can be applied to comparative modified numerals. With respect to UE *more than n*, the proposal of Hackl (2000, 2002) was sketched. Moreover, it was stated that a decompositional analysis of *fewer than*, as discussed at length in that section, can straightforwardly be applied to DE modified numerals of the form *fewer than n*. Below, in figures LI and LII an analysis of both types of quantifiers is sketched by means of example derivations that incorporate ideas of Büring (2007a), Hackl (2000, 2002) and others mentioned above.



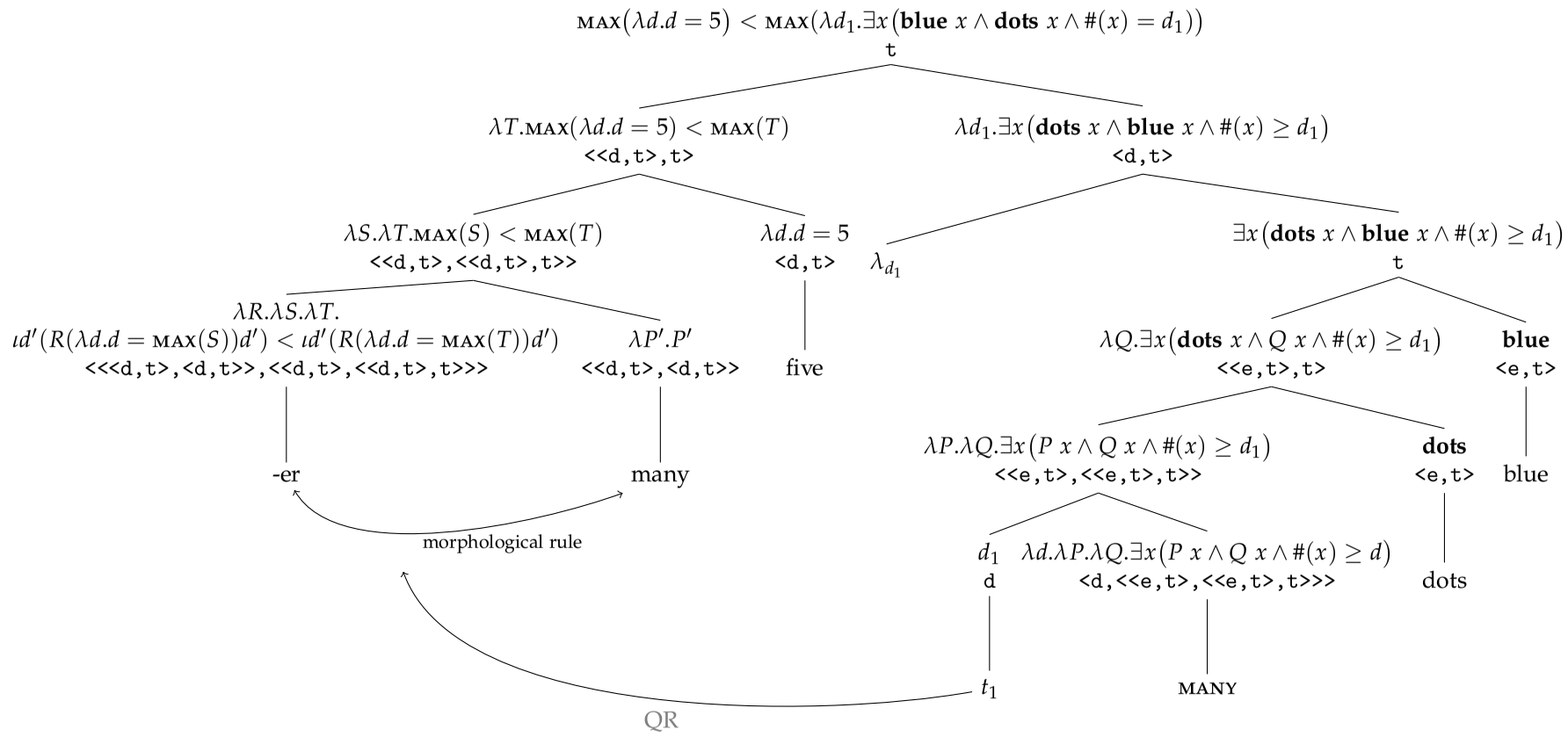


Figure LI. Derivation of *more than five*.

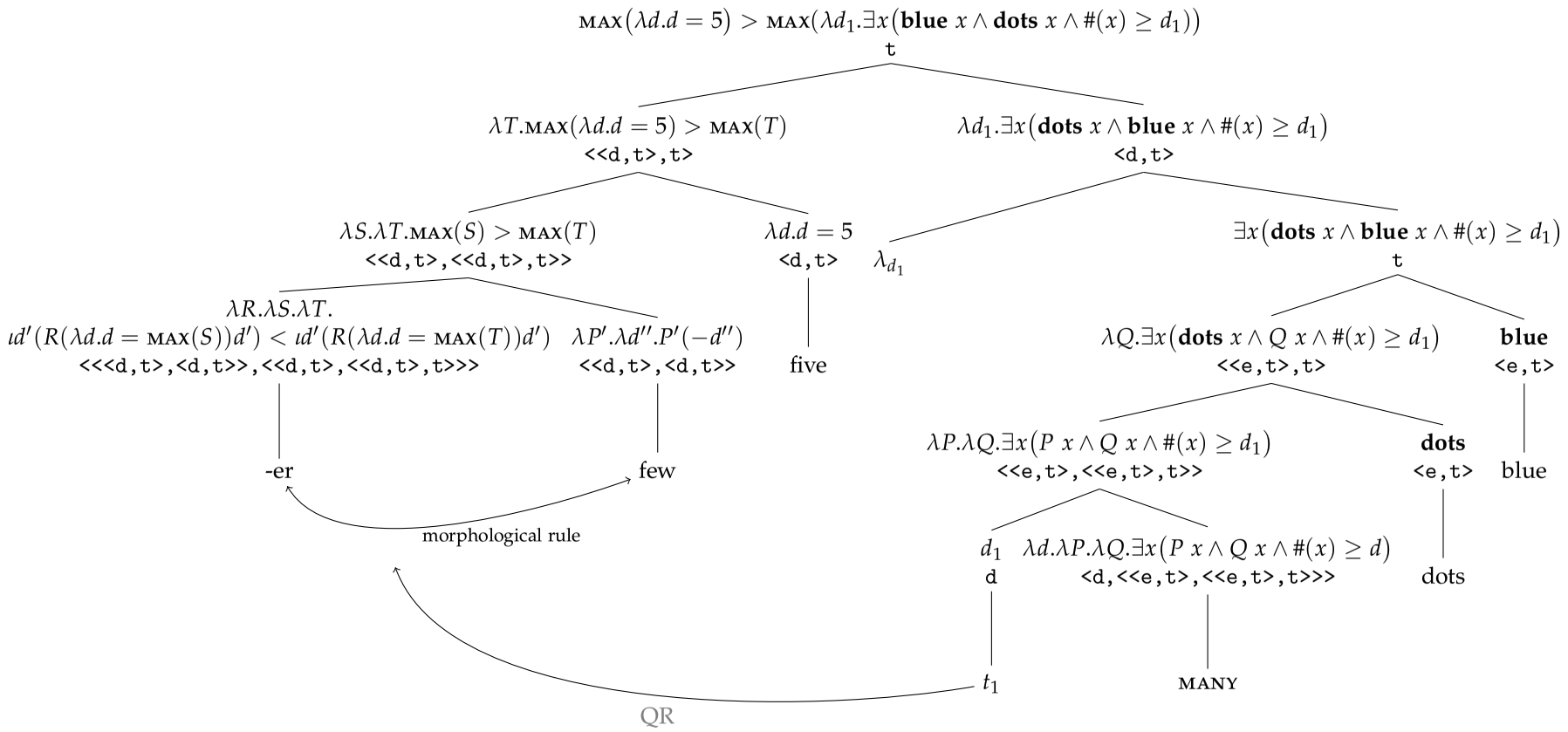


Figure LII. Derivation of *fewer than five*.

REFERENCES

Abels, K., & Martí, L. (2010). A unified approach to split scope. *Natural Language Semantics*, *18*(4), 435–470. doi: 10.1007/s11050-010-9060-8

Ades, A. E., & Steedman, M. J. (1982). On the order of words. *Linguistics and Philosophy*, *4*(4), 517–558. doi: 10.1007/bf00360804

Ajtai, M., & Fagin, R. (1990). Reachability is harder for directed than for undirected finite graphs. *The Journal of Symbolic Logic*, *55*(1), 113–150. doi: 10.1109/SFCS.1988.21952

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*(3), 191–238. doi: 10.1016/0010-0277(88)90020-0

Amann, H., & Escher, J. (2001). *Analysis III. Grundstudium Mathematik*. Basel: Birkhäuser Verlag.

Anderson, J. R. (1989). A rational analysis of human memory. In I. Roediger Henry L. & F. I. Craik (Eds.), *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving* (pp. 195–210). Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (2015). *Cognitive Psychology and its Implications* (8th ed.). New York, NY: Macmillan.

Ariel, M. (2004). Most. *Language*, *80*(4), 658–706.

Arora, S., & Barak, B. (2009). *Computational Complexity: A Modern Approach*. Cambridge University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47-89. doi: 10.1016/s0079-7421(08)60452-1

Bader, M., & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics*, *46*(02), 273–330. doi: 10.1017/S0022226709990260

Baggio, G., & van Lambalgen, M. (2007). The processing consequences of the imperfective paradox. *Journal of Semantics*, *24*(4), 307–330. doi: 10.1093/jos/ffm005

Balakrishnan, J. D., & Ashby, F. G. (1992). Subitizing: Magical numbers or mere superstition? *Psychological Research*, *54*(2), 80–90.

doi: 10.1007/bf00937136

Barker, C. (2005). Remark on Jacobson 1999: Crossover as a local constraint. *Linguistics and Philosophy*, *28*(4), 447–472. doi: 10.1007/s10988-004-5327-1

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. doi: 10.1016/j.jml.2012.11.001

Barton, G. E., Berwick, R. C., & Ristad, E. S. (1987). *Computational Complexity and Natural Language*. Cambridge, MA: MIT Press.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*(2), 159–219. doi: 10.1007/BF00350139

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *ArXiv preprints*. Retrieved from http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bech, G. (1955). *Studien über das deutsche Verbum infinitum*. Tübingen: Niemeyer. (Second unrevised edition published 1983)

Beck, S. (2001). Reciprocals are definites. *Natural Language Semantics*, *9*(1), 69–138. doi: 10.1023/a:1012203407127

Beck, S. (2011). Comparison constructions. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (pp. 1341–1389). Berlin: De Gruyter. doi: 10.1515/9783110255072.1341

Beck, S. (2012a). DegP scope revisited. *Natural Language Semantics*, *20*(3), 227–272. doi: 10.1007/s11050-012-9081-6

Beck, S. (2012b). Lucinda driving too fast again–the scalar properties of ambiguous than-clauses. *Journal of Semantics*, *30*(1), 1–63. doi: 10.1093/jos/ffr011

Bhatt, R., & Pancheva, R. (2004). Late merger of degree clauses. *Linguistic Inquiry*, *35*(1), 1-45. doi: 10.1162/002438904322793338

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. doi: 10.1037/0033-295X.113.4.700

Bott, O. (2010). *The Processing of Events*. Amsterdam: John Benjamins.

(Linguistics Today 162) doi: 10.1075/la.162

Bott, O., Featherston, S., Radó, J., & Stolterfoht, B. (2011). The application of experimental methods in semantics. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (pp. 305–321). Berlin: De Gruyter. doi: 10.1515/9783110226614.305

Bott, O., Klein, U., & Schlotterbeck, F. (2013). Witness sets, polarity reversal and the processing of quantified sentences. In M. Aloni, M. Franke, & F. Roelofsen (Eds.), *Proceedings of the 19th Amsterdam Colloquium.* Retrieved from http://www.illc.uva.nl/AC/AC2013/Proceedings

Bott, O., Klein, U., & Schlotterbeck, F. (2015). *Empty-set effects in the comprehension and verification of quantifiers.* (Talk presented at the "Experimental Approaches to Semantics Workshop" at ESSLLI 2015, Universitat Pompeu Fabra, 3–7 August 2015)

Bott, O., & Radó, J. (2009). How to provide exactly one interpretation for each sentence. In S. Featherston & S. Winkler (Eds.), *The Fruits of Empirical Linguistics* (pp. 25–46). Berlin: De Gruyter.

Bott, O., & Schlotterbeck, F. (2015). The processing domain of scope interaction. *Journal of Semantics*, *32*(1), 39–92. doi: 10.1093/jos/fft015

Bott, O., Schlotterbeck, F., & Klein, U. (n.d.). *Empty-set effects in quantifier interpretation.* (under review)

Bott, O., Schlotterbeck, F., & Szymanik, J. (2011). Interpreting tractable versus intractable reciprocal sentences. In *Proceedings of the Ninth International Workshop on Computational Semantics* (pp. 75–84). Retrieved from http://aclweb.org/anthology/W/W11/W11-0109.pdf

Bott, O., & Sternefeld, W. (2017). An event semantics with continuations for incremental interpretation. *Journal of Semantics*, *34*. doi: 10.1093/jos/ffw013

Brasoveanu, A., & Dotlačil, J. (2013). What a rational interpreter would do: Building, ranking, and updating quantifier scope representations in discourse. In M. Aloni, M. Franke, & F. Roelofsen (Eds.), *Proceedings of Amsterdam Colloquium 2013.* Retrieved from http://www.illc.uva.nl/AC/AC2013/Proceedings

Bresnan, J. W. (1973). Syntax of the comparative clause construction in English. *Linguistic Inquiry*, *4*(3), pp. 275-343.

Bueti, D., & Walsh, V. (2009). The parietal cortex and the representation of time, space, number and other magnitudes. *Philosophical*

Transactions of the Royal Society B: Biological Sciences, *364*(1525), 1831–1840. doi: 10.1098/rstb.2009.0028

Büring, D. (1997). The great scope inversion conspiracy. *Linguistics and Philosophy*, *20*(2), 175–194. doi: 10.1023/A:1005397026866

Büring, D. (2005). *Binding Theory*. Cambridge, MA: Cambridge University Press.

Büring, D. (2007a). Cross-polar nomalies. In *Proceeding of Semantics and Linguistic Theory XVII*. doi: 10.3765/salt.v17i0.2957

Büring, D. (2007b). More or less. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society* (Vol. 43, pp. 3–17).

Carnap, R. (1952). Meaning postulates. *Philosophical Studies*, *3*(5), 65–73. doi: 10.1007/BF02350366

Chemla, E., & Singh, R. (2014a). Remarks on the experimental turn in the study of scalar implicature, part I. *Language and Linguistics Compass*, *8*(9), 373–386. doi: 10.1111/lnc3.12081

Chemla, E., & Singh, R. (2014b). Remarks on the experimental turn in the study of scalar implicature, part II. *Language and Linguistics Compass*, *8*(9), 373–386. doi: 10.1111/lnc3.12081

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1977). On WH-movement. In P. W. Culicover, T. Wasow, & A. Akmajian (Eds.), *Formal Syntax* (p. 71-132). New York, NY: Academic Press.

Clark, H., & Chase, W. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472-517. doi: 10.1016/0010-0285(72)90019-9

Clark, R., & Grossman, M. (2007). Number sense and quantifier interpretation. *Topoi*, *26*(1), 51–62. doi: 10.1007/s11245-006-9008-2

Clifton, C., Staub, A., & Rayner, K. (2004). Eye movements in reading words and sentences. In R. P. G. van Gompel (Ed.), *Eye Movements: A Window on Mind and Brain* (pp. 341–372). Amsterdam: Elsevier.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, *42*(2-3), 393405. doi: 10.1016/0004-3702(90)90060-d

Corver, N., & Zwarts, J. (2006). Prepositional numerals. *Lingua*, *116*(6), 811–835. doi: 10.1016/j.lingua.2005.03.008

Cresswell, M. J. (1976). The semantics of degree. In B. Partee (Ed.), *Montague Grammar* (pp. 261–292). New York, NY: Academic

Press.

Cummins, C., & Katsos, N. (2010). Comparative and superlative quantifiers: Pragmatic effects of comparison type. *Journal of Semantics*, *27*(3), 271–305. doi: 10.1093/jos/ffq006

Dalrymple, M., Kanazawa, M., Kim, Y., McHombo, S., & Peters, S. (1998). Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy*, *21*(2), 159–210. doi: 10.1023/a:1005330227480

Dayan, P., & Abbott, L. F. (2001). *Theoretical Neuroscience* (Vol. 806). Cambridge, MA: MIT Press.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. Oxford: 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. Rosetti, & M. Kawato (Eds.), *Sensorimotor Foundations of Higher Cognition* (p. 527-574). Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199231447.001.0001

Dehaene, S., & Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *J. Cognitive Neuroscience*, *5*(4), 390–407. doi: 10.1162/jocn.1993.5.4.390

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, *56*(2), 384–398. doi: 10.1016/j.neuron.2007.10.004

Dekking, F. M. (2005). *A Modern Introduction to Probability and Statistics: Understanding why and how*. New York, NY: Springer.

Demberg, V. (2012). Incremental derivations in CCG. In *Proceedings of the 11th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+ 11)* (pp. 198–206). Retrieved from https://aclweb.org/anthology/W/W12/W12-4600.pdf

Deschamps, I., Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers and numerosity perception. *Cognition*, *143*, 115–128. doi: 10.1016/j.cognition.2015.06.006

de Swart, H. (2000). Scope ambiguities with negative quantifiers. In H. von Heusinger & U. Egli (Eds.), *Reference and Anaphoric Relations* (pp. 109–132). Dordrecht: Springer. doi: 10.1007/978-94-011-3947-2˙6

Dosher, B. A. (1976). The retrieval of sentences from memory: A speed-accuracy study. *Cognitive psychology*, *8*(3), 291–310. doi: 10.1016/0010-0285(76)90009-8

Dosher, B. A. (1979). Empirical approaches to information pro-

cessing: Speed-accuracy tradeoff functions or reaction time – a reply. *Acta Psychologica*, *43*(5), 347–359. doi: 10.1016/0001-6918(79)90029-5

Dotlačil, J., Szymanik, J., & Zajenkowski, M. (2014). Probabilistic semantic automata in the verification of quantified statements. In P. Bello, M. McShane, M. Guarini, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1778–1783). Retrieved from https://mindmodeling.org/cogsci2014

Fauconnier, G. (1978). Implication reversal in a natural language. In F. Guenthner & S. J. Schmidt (Eds.), *Formal Semantics and Pragmatics for Natural Languages* (pp. 289–301). Dordrecht: Springer.

Featherston, S. (2007). Data in generative grammar: The stick and the carrot. *Theoretical Linguistics*, *33*(3), 269–318. doi: 10.1515/TL.2007.020

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. doi: 10.1016/j.tics.2004.05.002

Filik, R., Paterson, K. B., & Liversedge, S. P. (2004). Processing doubly quantified sentences: Evidence from eye movements. *Psychonomic Bulletin & Review*, *11*(5), 953-959. doi: 10.3758/BF03196727

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. doi: 10.1111/j.1469-8986.1983.tb00920.x

Fodor, J. A. (1986). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Forstmann, B., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*(1), 641–666. doi: 10.1146/annurev-psych-122414-033645

Fox, D., & Hackl, M. (2007). The universal density of measurement. *Linguistics and Philosophy*, *29*(5), 537–586. doi: 10.1007/s10988-006-9004-4

Freunberger, D., & Nieuwland, M. S. (2016). Incremental comprehension of spoken quantifier sentences: Evidence from brain potentials. *Brain Research*, *1646*, 475–481. doi: 10.1016/j.brainres.2016.06.035

Frixione, M. (2001). Tractable competence. *Minds and Machines*, *11*(3),

379–397. doi: 10.1023/A:1017503201702

Gallistel, C., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*(1-2), 43–74. doi: 10.1016/0010-0277(92)90050-R

Gamut, L. (1991a). *Logic, Language, and Meaning* (Vol. 1). Chicago: University of Chicago Press.

Gamut, L. (1991b). *Logic, Language, and Meaning* (Vol. 2). Chicago: University of Chicago Press.

Garey, M. R., & Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Bell Telephone Laboratories.

Gelman, A., & Hill, J. (2007). *Data Aanalysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Geurts, B. (1996). On no. *Journal of Semantics*, *13*(1), 67–86. doi: 10.1093/jos/13.1.67

Geurts, B., Katsos, N., Moons, J., & Noordman, L. (2010). Scalar quantifiers: Logic, acquisition, and processing. *Language and Cognitive Processes*, *25*, 130–148. doi: 10.1080/01690960902955010

Geurts, B., & Nouwen, R. (2007). "At least" et al.: the semantics of scalar modifiers. *Language*, *83*, 533–559.

Geurts, B., & van der Slik, F. (2005). Monotonicity and processing load. *Journal of Semantics*, *22*(1), 97-117. doi: 10.1093/jos/ffh018

Giannakidou, A. (2011). Negative and positive polarity items: Variation, licensing, and compositionality. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (pp. 1660–1712). Berlin: De Gruyter. doi: 10.1515/9783110255072.1660

Gibson, E., Jacobson, P., Graff, P., Mahowald, K., Fedorenko, E., & Piantadosi, S. T. (2015). A pragmatic account of complexity in definite antecedent-contained-deletion relative clauses. *Journal of Semantics*, *32*(4), 579–618. doi: 10.1093/jos/ffu006

Gierasimczuk, N., & Szymanik, J. (2009). Branching quantification v. two-way quantification. *Journal of Semantics*, *26*(4), 367–392. doi: 10.1093/jos/ffp008

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, *5*(1), 10–16. doi: 10.1016/S1364-6613(00)01567-9

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions,

and reward. *Neuron*, *36*(2), 299–308. doi: 10.1016/S0896-6273(02)00971-6

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin & C. Fox (Eds.), *The Handbook of Contemporary Semantic Theory* (Second ed.). Hoboken, NJ: Wiley-Blackwell.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York, NY: Wiley.

Grosu, A., & Horvath, J. (2006). Reply to Bhatt and Pancheva's "late merger of degree clauses": The irrelevance of (non)conservativity. *Linguistic Inquiry*, *37*(3), 457–483. doi: 10.1162/ling.2006.37.3.457

Hackl, M. (2000). *Comparative Quantifiers* (Doctoral dissertation, Massachusetts Institute of Technology). Retrieved from http://web.mit.edu/hackl/www/papers/files/NThesis5.pdf

Hackl, M. (2002). Comparative quantifiers and plural predication. In K. Megerdoomian & L. A. Barel (Eds.), *Proceedings of the 20th West Coast Conference on Formal Linguistics* (pp. 234–247). Somerville, MA: Cascadilla Press.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: *most* versus *more than half*. *Natural Language Semantics*, *17*(1), 63–98. doi: 10.1007/s11050-008-9039-x

Hackl, M., Koster-Hale, J., & Varvoutis, J. (2012). Quantification and ACD: Evidence from real-time sentence processing. *Journal of Semantics*, *29*(2), 145–206. doi: 10.1093/jos/ffr009

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, *17*(7), 572–576. doi: 10.1111/j.1467-9280.2006.01746.x

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, *10*(9), 397–412. doi: 10.1111/lnc3.12196

Halpern, J. Y. (2003). *Reasoning about uncertainty*. MIT Press.

Hamm, F. (1989). *Natürlich-sprachliche Quantoren: Modelltheoretische Untersuchungen zu universellen semantischen Beschränkungen*. Berlin: De Gruyter.

Hamm, F., Kamp, H., & Van Lambalgen, M. (2006). There is no opposition between formal and cognitive semantics. *Theoretical Linguistics*, *32*(1), 1. doi: 10.1515/TL.2006.001

Hamm, F., & Zimmermann, T. E. (2002). Quantifiers and anaphora. In F. Hamm & T. E. Zimmermann (Eds.), *Semantics* (pp. 137–172). Hamburg: Buske.

Heim, I. (2000). Degree operators and scope. In *Proceedings of Semantics and Linguistic Theory X* (pp. 40–61). Linguistic Society of America. doi: 10.3765/salt.v10i0.3102

Heim, I. (2006). "Little". In *Proceedings of Semantics and Linguistic Theory XVI* (pp. 35–58). Linguistic Society of America. doi: 10.3765/salt.v16i0.2941

Heim, I. (2008). Decomposing antonyms? In A. Grønn (Ed.), *Proceedings of SuB12* (pp. 212–225). Retrieved from http://www.hf.uio.no/ilos/forskning/aktuelt/arrangementer/konferanser/2007/SuB12/proceedings/

Heim, I., & Kratzer, A. (1998). *Semantics in Generative Grammar*. Malden, MA: Wiley-Blackwell.

Heim, I., Lasnik, H., & May, R. (1991). On "reciprocal scope". *Linguistic Inquiry*, *22*(1), 173–192.

Heim, S., Amuntsa, K., Drai, D., Eickhoff, S. B., Hautvast, S., & Grodzinsky, Y. (2012). The language-number interface in the brain: A complex parametric study of quantifiers and quantities. *Frontiers in Evolutionary Neurosciences*, *4*(4), 1–12. doi: 10.3389/fnevo.2012.00004

Hella, L., Väänänen, J., & Westerståhl, D. (1997). Definability of polyadic lifts of generalized quantifiers. *Journal of Logic, Language and Information*, *6*(3), 305–335. doi: 10.1023/A:1008215718090

Hintikka, J. (1973). Quantifiers vs. quantification theory. *Dialectica*, *27*(3), 329–358. doi: 10.1111/j.1746-8361.1973.tb00624.x

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to Automata Theory, Languages, and Computation*. Reading: Addison Wesley.

Hubbard, E. M., Piazza, M., Pinel, P., & Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature Reviews Neuroscience*(6), 435–448. doi: 10.1038/nrn1684

Hunter, T., Lidz, J., Odi, D., & Wellwood, A. (2016). On how verification tasks are related to verification procedures: A reply to Kotek et al. *Natural Language Semantics*. (online first) doi: 10.1007/s11050-016-9130-7

Immerman, N. (1999). *Descriptive Complexity*. New York: Springer. (Series: Graduate Texts in Computer Science) doi: 10.1007/978-1-4612-0539-5

Isaac, A., Szymanik, J., & Verbrugge, R. (2014). Logic and complexity in cognitive science. In A. Baltag & S. Smets (Eds.), *Johan van Benthem on Logic and Information Dynamics* (Vol. 5,

p. 787-824). Cham (ZG): Springer International Publishing. doi: 10.1007/978-3-319-06025-5˙30

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*(3), 1221–1247. doi: 10.1016/j.cognition.2007.06.004

Jacob, S. N., Vallentin, D., & Nieder, A. (2012). Relating magnitudes: The brain's code for proportions. *Trends in Cognitive Sciences*, *16*(3), 157–166. doi: 10.1016/j.tics.2012.02.002

Jacobs, J. (1980). Lexical decomposition in Montague-grammar. *Theoretical Linguistics*, *7*(1), 121–136. doi: 10.1515/thli.1980.7.1-3.121

Jacobs, J. (1982). *Syntax und Semantik der Negation im Deutschen*. München: Fink.

Jäger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. doi: 10.1016/j.jml.2007.11.007

Joseph, D. A., & Plantings, W. H. (1985). On the complexity of reachability and motion planning questions (extended abstract). In *Proceedings of the first Annual Symposium on Computational Geometry - SCG 85* (pp. 62–66). Association for Computing Machinery (ACM). doi: 10.1145/323233.323242

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behaviour*, *10*(3), 244–253. doi: 10.1016/S0022-5371(71)80051-8

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238.

Kamp, H., & Reyle, U. (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Springer.

Kanazawa, M. (2013). Monadic quantifiers recognized by deterministic pushdown automata. In M. F. Maria Aloni & F. Roelofsen (Eds.), *Proceedings of the 19th Amsterdam Colloquium* (pp. 139–146). Retrieved from http://www.illc.uva.nl/AC/AC2013/Proceedings

Kannan, D. (1979). *An Introduction to Stochastic Processes*. New York: North Holland.

Karp, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of Computer Computations* (pp. 85–103). Dordrecht: Springer. doi: 10.1007/978-1-4684-2001-2˙9

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*(4), 498–525. doi: 10.2307/1418556

Kaup, B., Zwaan, R., & Lüdtke, J. (2007). The experiential view of language comprehension: How is negated text information represented? In F. Schmalhofer & C. Perfetti (Eds.), *Higher Level Language Processes in the Brain: Inference and Comprehension Processes* (p. 255 - 288). Mahwah, NJ: Erlbaum.

Keenan, E. L. (2006). Quantifiers: Semantics. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (Second ed., Vol. 10, pp. 302–308). Oxford: Elsevier.

Keenan, E. L., & Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy*, *9*(3), 253–326. doi: 10.1007/BF00630273

Kennedy, C. (1997). *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison* (Doctoral dissertation, University of California, Santa Cruz). Retrieved from http://semantics.uchicago.edu/kennedy/docs/ck-thesis.pdf

Kennedy, C. (2001). Polar opposition and the ontology of degrees. *Linguistics and Philosophy*, *24*(1), 33–70. doi: 10.1023/A:1005668525906

Kennedy, C. (2002). Comparative deletion and optimality in syntax. *Natural Language & Linguistic Theory*, *20*(3), 553-621. doi: 10.1023/A:1015889823361

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*(1), 1–45. doi: 10.1007/s10988-006-9008-0

Kennedy, C. (2015). A de-Fregean semantics (and neo-Gricean pragmatics) for modified and unmodified numerals. *Semantics and Pragmatics*, *8*(10), 1–44. doi: 10.3765/sp.8.10

Kennedy, C., & McNally, L. (2005a). Scale structure, degree modification and the semantics of gradable predicates. *Language*, *81*(2), 345–381. doi: 10.1007/s11050-009-9045-7

Kennedy, C., & McNally, L. (2005b). The syntax and semantics of multiple degree modification in English. In S. Müller (Ed.), *Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar* (pp. 178–91). CSLI Publications. Retrieved from http://web.stanford.edu/group/cslipublications/cslipublications/HPSG/2005/

Kerem, N., Friedmann, N., & Winter, Y. (2011). Typicality effects and

the logic of reciprocity. In E. Cormany, S. Ito, & D. Lutz (Eds.), *Proceedings of Semantics and Linguistic Theory XIX* (pp. 257–274). eLanguage. doi: 10.3765/salt.v19i0.2537

Kirousis, L. M., & Kolaitis, P. G. (2001). On the complexity of model checking and inference in minimal models. In *Proceedings of the 6th International Conference on Logic Programming and Nonmonotonic Reasoning* (pp. 42–53). London: Springer.

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, *4*(1), 1–45. doi: 10.1007/bf00351812

Kontinen, J., & Szymanik, J. (2008). A remark on collective quantification. *Journal of Logic, Language and Information*, *17*(2), 131–140. doi: 10.1007/s10849-007-9055-0

Koster-Moeller, J., Varvoutis, J., & Hackl, M. (2008). Verification procedures for modified numeral quantifiers. In N. Abner & J. Bishop (Eds.), *Proceedings of the 27th West Coast Conference on Formal Linguistics.* Somerville, MA: Cascadilla Proceedings Project.

Kotek, H., Sudo, Y., & Hackl, M. (2015). Experimental investigations of ambiguity: The case of 'most'. *Natural Language Semantics*, *23*(2), 119–156. doi: 10.1007/s11050-015-9113-0

Kotek, H., Sudo, Y., Hackl, M., & Howard, E. (2011). Most meanings are superlative. In J. Runner (Ed.), *Experiments at the Interfaces* (pp. 101–145). Brill Academic Publishers. doi: 10.1163/9781780523750˙005

Kotek, H., Sudo, Y., Howard, E., & Hackl, M. (2011). Three readings of 'most'. In *Proceedings of Semantics and Linguistic Theory XXI* (pp. 353–372). Linguistic Society of America. doi: 10.3765/salt.v21i0.2621

Krifka, M. (1999). At least some determiners aren't determiners. In K. Turner (Ed.), *The Semantics/Pragmatics Interface From Different Points of View* (pp. 257–292). Oxford: Elsevier Science.

Krifka, M. (2011). Varieties of semantic evidence. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (pp. 242–267). Berlin: De Gruyter. doi: 10.1515/9783110226614.242

Ladusaw, W. (1980). *Polarity Sensitivity as Inherent Scope Relations*. New York: Garland Pub. (Series: Outstanding Dissertations in Linguistics)

Langendoen, D. T. (1978). The logic of reciprocity. *Linguistic Inquiry*, *9*(2), 177–197.

Lassiter, D. (2015). Adjectival modification and gradation. In S. Lap-

pin & C. Fox (Eds.), *The Handbook of Contemporary Semantic Theory* (Second ed., pp. 141–167). Hoboken, NJ: Wiley-Blackwell. doi: 10.1002/9781118882139.ch5

Lawler, G. (1995). *Introduction to Stochastic Processes*. New York: Chapman & Hall.

Lechner, W. (2001). Reduced and phrasal comparatives. *Natural Language & Linguistic Theory*, *19*(4), 683–735.

Levy, R. (2014). Using R formulae to test for main effects in the presence of higher-order interactions. *ArXiv e-prints*. Retrieved from http://arxiv.org/abs/1405.2094

Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*, *44*(1), 27–46. doi: 10.1007/s10936-014-9329-z

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of 'most'. *Natural Language Semantics*, *19*(3), 227–256. doi: 10.1007/s11050-010-9062-6

Lindström, P. (1966). First order predicate logic with generalized quantifiers. *Theoria*, *32*(3), 186–195. doi: 10.1111/j.1755-2567.1966.tb00600.x

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511Mckay. doi: 10.1037/a0022643

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press. Retrieved from http://www.inference.phy.cam.ac.uk/mackay/itila/

Mallot, H. A. (2013). *Computational Neuroscience*. Cham (ZG): Springer. doi: 10.1007/978-3-319-00861-5

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: Henry Holt and Co., Inc.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*(4), 287–330. doi: 10.1037/0033-295X.86.4.287

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General*, *118*(4), 346–373. doi: 10.1037//0096-3445.118.4.346

McMillan, C. T., Clark, R., Moore, P., Devita, C., & Gross-

man, M. (2005). Neural basis for generalized quantifier comprehension. *Neuropsychologia*, *43*(12), 1729–1737. doi: 10.1016/j.neuropsychologia.2005.02.012

McMillan, C. T., Clark, R., Moore, P., & Grossman, M. (2006). Quantifier comprehension in corticobasal degeneration. *Brain and Cognition*, *62*(3), 250–260. doi: 10.1016/j.bandc.2006.06.005

Montague, R. (1973). The proper treatment of quantification in ordinary english. In K. Hintikka, J. Moravcski, & S. P. (Eds.), *Approaches to Natural Language* (pp. 221–242). Dordrecht: Reidel.

Moschovakis, Y. N. (1994). Sense and denotation as algorithm and value. In J. Oikkonen & Väänänen (Eds.), *Logic Colloquium '90: ASL Summer Meeting in Helsinski* (Vol. 2, pp. 210–249). Heidelberg: Springer.

Mostowski, M. (1998). Computational semantics for monadic quantifiers. *Journal of Applied Non-Classical Logics*, *8*(1-2), 107-121. doi: 10.1080/11663081.1998.10510934

Mostowski, M., & Szymanik, J. (2007). Computational complexity of some ramsey quantifiers in finite models. *The Bulletin of Symbolic Logic*, *13*, 281–282. doi: 10.1007/978-3-319-28749-2˙7

Mostowski, M., & Szymanik, J. (2012). Semantic bounds for everyday language. *Semiotica*, *188*(1), 363–372. doi: 10.1515/sem-2012-0022

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*(5109), 1519–1520. doi: 10.1038/2151519a0

Neeleman, A., Van de Koot, H., & Doetjes, J. (2004). Degree expressions. *Linguistic Review*, *21*(1), 1–66. doi: 10.1515/tlir.2004.001

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annu. Rev. Neurosci.*, *32*(1), 185–208. doi: 10.1146/annurev.neuro.051508.135550

Nieder, A., & Merten, K. (2007). A labeled-line code for small and large numerosities in the monkey prefrontal cortex. *The Journal of neuroscience*, *27*(22), 5986–5993. doi: 10.1523/JNEUROSCI.1056-07.2007

Nieder, A., & Miller, E. K. (2004). A parieto-frontal network for visual numerical information in the monkey. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(19), 7457-7462. doi: 10.1073/pnas.0402239101

Nieuwland, M. S. (2016). Quantification, prediction, and the online impact of sentence truth-value: Evidence from event-related po-

tentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 316–334. doi: 10.1037/xlm0000173

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. doi: 10.1111/j.1467-9280.2008.02226.x

Nouwen, R. (2010a). Two kinds of modified numerals. *Semantics and Pragmatics*, *3*. (article 3) doi: 10.3765/sp.3.3

Nouwen, R. (2010b). What's in a quantifier? In M. Everaert, T. Lentz, H. de Mulder, O. Nilsen, & A. Zondervan (Eds.), *The Linguistic Enterprise*. Amsterdam: John Benjamins.

Nouwen, R. (2015). Modified numerals: The epistemic effect. In L. lonso Ovalle & P. Menendez-Benito (Eds.), *Epistemic Indefinites*. Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, *32*(01), 69–84. doi: 10.1017/s0140525x09000284

Partee, B. (1989). Many quantifiers. In J. Powers & K. de Jong (Eds.), *Proceedings of the 5th Eastern States Conference on Linguistics* (pp. 383–402).

Penka, D. (2011). *Negative Indefinites*. Oxford: Oxford University Press. (Series: Oxford Studies in Theoretical Linguistics, vol.: 32) doi: 10.1093/acprof:oso/9780199567263.001.0001

Penka, D. (2012). Split scope of negative indefinites. *Language and Linguistics Compass*, *6*(8), 517–532. doi: 10.1002/lnc3.349

Penka, D. (2015). At most at last. In E. Csipak & H. Zeijlstra (Eds.), *Proceedings of SuB19* (pp. 463–480).

Penka, D., & Stechow, A. v. (2001). Negative Indefinita unter Modalverben. *Linguistische Berichte*, *9*, 263–286.

Peters, S., & Westerståhl, D. (2006). *Quantifiers in Language and Logic*. New York: Oxford University Press.

Piantadosi, S. T. (2011). *Learning and the Language of Thought* (Doctoral dissertation, Massachusetts Institute of Technology). Retrieved from https://colala.bcs.rochester.edu/papers/piantadosi_thesis.pdf

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547–555. doi: 10.1016/j.neuron.2004.10.014

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approx-

imate arithmetic in an Amazonian indigene group. *Science*, *306*, 499–503. doi: 10.1126/science.1102085

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of 'most': Semantics, numerosity, and psychology. *Mind and Language*, *24*(5), 554–585. doi: 10.1111/j.1468-0017.2009.01374.x

Platt, M., Dayan, P., Dehaene, S., McCabe, K., Menzel, R., Phelps, E., ... Singer, W. (2008). Neuronal correlates of decision making. In C. Engel & W. Singer (Eds.), *Better Than Conscious? Decision Making, the Human Mind, and Implications for Institutions* (pp. 125–154). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262195805.003.0006

Poeppel, D., & Embick, D. (2005). Defining the relation between linguistics and neuroscience. In A. Cutler (Ed.), *Twenty-first Century Psycholinguistics: Four Cornerstones* (pp. 103–118). Mahwah, NJ: Erlbaum.

Poor, H. V. (1994). *An Introduction to Signal Detection and Estimation*. New York, NY: Springer. doi: 10.1007/978-1-4757-2341-0

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience*, *26*(1), 381–410. doi: 10.1146/annurev.neuro.26.041002.131112

Pouget, A., Deneve, S., & Duhamel, J.-R. (2002). A computational perspective on the neural basis of multisensory spatial representations. *Nature Reviews Neuroscience*, *3*(9), 741–747. doi: 10.1038/nrn914

Pouget, A., & Sejnowski, T. J. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, *9*(2), 222–237. doi: 10.1162/jocn.1997.9.2.222

Pratt-Hartmann, I. (2004). Fragments of language. *Journal of Logic, Language and Information*, *13*(2), 207–223. doi: 10.1023/b:jlli.0000024735.97006.5a

Pratt-Hartmann, I. (2008). On the computational complexity of the numerically definite syllogistic and related logics. *Bulletin of Symbolic Logic*, *14*(1), 1–28. doi: 10.2178/bsl/1208358842

Pratt-Hartmann, I. (2010). Computational complexity in natural language. In A. Clark, C. Fox, & S. Lappin (Eds.), *The Handbook of Computational Linguistics and Natural Language Processing*. Hoboken, NJ: Wiley-Blackwell. (Series: Blackwell Handbooks in Linguistics, vol.: 57)

Pratt-Hartmann, I., & Third, A. (2006). More fragments of language. *Notre Dame Journal of Formal Logic*, *47*(2), 151–177.

Pylkkänen, L., Brennan, J., & Bemis, D. K. (2011). Grounding the cognitive neuroscience of semantics in linguistic theory. *Language and Cognitive Processes*, *26*(9), 1317–1337.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology*, *53*(3), 195–237. doi: 10.1016/j.cogpsych.2005.10.002

Ratcliff, R. (2008). Modeling aging effects on two-choice tasks: Response signal and response time data. *Psychology and Aging*, *23*(4), 900–916. doi: 10.1037/a0013930

Rautenberg, W. (2009). *A Concise Introduction to Mathematical Logic*. New York: Springer. doi: 10.1007/978-1-4419-1221-3

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. doi: 10.1037//0033-2909.124.3.372

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science*, *181*(4099), 574–576. doi: 10.1126/science.181.4099.574

Reif, J. H. (1985). *Complexity of the Generalized Mover's Problem* (Tech. Rep.). Center for Research in Computing Technology. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA161378

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science*, *19*(6), 607–614. doi: 10.1111/j.1467-9280.2008.02130.x

Ristad, E. S. (1993). *The Language Complexity Game*. Cambridge: MIT Press.

Romoli, J. (2015). A structural account of conservativity. *Semantics-Syntax Interface*, *2*(1), 28–57. Retrieved from http://semantics-syntax.ut.ac.ir/

Rullmann, H. (1995). *Maximality in the Semantics of Wh-Constructions* (Doctoral dissertation, University of Massachusetts at Amherst). Retrieved from http://scholarworks.umass.edu/dissertations/AAI9524743

Ruys, E., & Winter, Y. (2011). Quantifier scope in formal linguistics. In D. M. Gabbay & F. Guenthner (Eds.), *Handbook of Philosophical Logic* (Vol. 16, pp. 159–225). Springer Netherlands. doi: 10.1007/978-94-007-0479-4˙3

Sabato, S., & Winter, Y. (2012). Relational domains and the interpretation of reciprocals. *Linguistics and Philosophy*, *35*(3), 191–241. doi: 10.1007/s10988-012-9117-x

Schlotterbeck, F. (2015). *The Process of Evaluating Comparative Quantifiers.* (Talk presented at the "Experimental Approaches to Semantics Workshop" at ESSLLI 2015, Universitat Pompeu Fabra, 3–7 August 2015)

Schlotterbeck, F., & Bott, O. (2013). Easy solutions for a hard problem? The computational complexity of reciprocals with quantificational antecedents. *Journal of Logic, Language and Information*, *22*(4), 363-390. doi: 10.1007/s10849-013-9181-9

Schöller, A., & Franke, M. (2015). Semantic values as latent parameters: Surprising few & many. In *Proceedings of Semantics and Linguistic Theory XV* (Vol. 25, pp. 143–162). doi: 10.3765/salt

Schöller, A., & Franke, M. (2016). How many manys? Exploring semantic theories with data-driven computational models. In N. Bade, P. Berezovskaya, & A. Schöller (Eds.), *Proceeding of SuB20* (pp. 622–639). Retrieved from http://www.sfs.uni-tuebingen.de/~mfranke/Papers/SchollerFranke_2016_How_many_manys.pdf

Schwarz, B. (2013). 'At least' and quantity implicature: Choices and consequences. In M. Aloni, M. Franke, & F. Roelofsen (Eds.), *Proceedings of the 19th Amsterdam Colloquium.*

Schwarzchild, R., & Wilkinson, K. (2002). Quantifiers in comparatives: A semantics of degree based on intervals. *Natural Language Semantics*, *10*(1), 1–41. doi: 10.1023/A:1015545424775

Scontras, G., Graff, P., & Goodman, N. D. (2012). Comparing pluralities. *Cognition*, *123*(1), 190–197. doi: 10.1016/j.cognition.2011.12.012

Seuren, P. (1979). Meer over minder dan hoeft. *De Nieuwe Taalgids*, *72*, 236–239.

Sevenster, M. (2006). *Branches of Imperfect Information: Logic, Games, and Computation* (Doctoral dissertation, ILLC, Amsterdam). Retrieved from https://www.illc.uva.nl/Research/Publications/Dissertations/DS-2006-06.text.pdf

Sherman, M. A. (1976). Adjectival negation and the comprehension of

multiply negated sentences. *Journal of Verbal Learning and Verbal Behaviour*, *15*, 143–157. doi: 10.1016/0022-5371(76)90015-3

Snippe, H. P. (1996). Parameter extraction from population codes: A critical assessment. *Neural Computation*, *8*(3), 511–529. doi: 10.1162/neco.1996.8.3.511

Solt, S. (2009). *The Semantics of Adjectives of Quantity* (Doctoral dissertation, The City University of New York). Retrieved from http://gradworks.umi.com/33/49/3349494.html

Solt, S. (2014). Q-adjectives and the semantics of quantity. *Journal of Semantics*, *32*(2), 221–273. doi: 10.1093/jos/ffto18

Solt, S. (2015). Measurement scales in natural language. *Language and Linguistics Compass*, *9*(1), 14–32. doi: 10.1111/lnc3.12101

Solt, S. (2016a). On measurement and quantification: The case of most and more than half. *Language*, *92*(1), 65–100. doi: 10.1353/lan.2016.0016

Solt, S. (2016b). *Proportional comparatives and relative scales.* Retrieved from http://www.iatl.org.il/wp-content/uploads/2016/07/iatl32_solt.pdf (Talk given at the 32nd annual meeting of the Israel Association for Theoretical Linguistics at Hebrew University of Jerusalem, 25 October 2016.)

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, *115*(11), 1497–1524. doi: 10.1016/j.lingua.2004.07.002

Steedman, M. (2000). *The Syntactic Process*. Cambridge, MA: MIT Press.

Steedman, M., & Baldridge, J. (2011). Combinatory categorial grammar. In R. Borsley & K. Borjars (Eds.), *Non-Transformational Syntax: Formal and Explicit Models of Grammar.* Oxford: Wiley-Blackwell.

Steinert-Threlkeld, S. (2016). Some properties of iterated languages. *Journal of Logic, Language and Information*, *25*(2), 191–213. doi: 10.1007/s10849-016-9239-6

Steinert-Threlkeld, S., & Icard, I., Thomas F. (2013). Iterating semantic automata. *Linguistics and Philosophy*, *36*(2), 151–173. doi: 10.1007/s10988-013-9132-6

Steinert-Threlkeld, S., Munneke, G.-J., & Szymanik, J. (2015). Alternative representations in formal semantics: A case study of quantifiers. In T. Brochhagen & F. R. N. Theiler (Eds.), *Proceedings of the 20th Amsterdam Colloquium* (pp. 368–377). Retrieved from http://semanticsarchive.net/Archive/mVkOTk2N/AC2015-proceedings.pdf

Sternberg, S. (1969). The discovery of processing stages: Extensions of donders' method. *Acta Psychologica*, *30*, 276–315. doi: 10.1016/0001-6918(69)90055-9

Sternefeld, W. (1998). Reciprocity and cumulative predication. *Natural Language Semantics*, *6*(3), 303–337. doi: 10.1023/a:1008352502939

Sternefeld, W. (2015). *When a Man Loves a Woman...* Retrieved from http://www.s395910558.online.de/Downloads/Quantification.pdf (submitted to "The Blackwell Companion to Semantics")

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260. doi: 10.1007/BF02289729

Svenonius, P., & Kennedy, C. (2006). Northern Norwegian degree questions and the syntax of measurement. In M. Frascarelli (Ed.), *Phases of Interpretation* (pp. 133–161). Berlin: De Gruyter.

Szabolcsi, A. (1987). Bound variables in syntax (are there any?). In J. Groenendijk, F. Veltman, & M. Stokhof (Eds.), *Sixth Amsterdam Colloquium Proceedings*. University of Amsterdam.

Szabolcsi, A. (2010). *Quantification*. Cambridge: Cambridge University Press.

Szabolcsi, A. (2014). Quantification and ACD: What is the evidence from real-time processing evidence for? A response to Hackl et al. (2012). *Journal of Semantics*, *31*, 135–145. doi: 10.1093/jos/ffs025

Szymanik, J. (2007). A comment on a neuroimaging study of natural language quantifier comprehension. *Neuropsychologia*, *45*(9), 2158–2160. doi: 10.1016/j.neuropsychologia.2007.01.016

Szymanik, J. (2009). *Quantifiers in TIME and SPACE - Computational Complexity of Generalized Quantifiers in Natural Language* (Doctoral dissertation, ILLC, Amsterdam). Retrieved from http://www.illc.uva.nl/Research/Publications/Dissertations/DS-2009-01.text.pdf

Szymanik, J. (2010). Computational complexity of polyadic lifts of generalized quantifiers in natural language. *Linguistics and Philosophy*, *33*(3), 215–250. doi: 10.1007/s10988-010-9076-z

Szymanik, J. (2016). *Quantifiers and Cognition. Logical and Computational Perspectives*. Cham (ZG): Springer. (Series: Studies in Linguistics and Philosophy, vol.: 96) doi: 10.1007/978-3-319-28749-2

Szymanik, J., & Thorne, C. (2015). Semantic complexity of quantifiers

and their distribution in corpora. In *Proceedings of the 11th International Conference on Computational Semantics* (pp. 64–69). London, UK: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W/W15/W15-0109

Szymanik, J., & Zajenkowski, M. (2010). Comprehension of simple quantifiers empirical evaluation of a computational model. *Cognitive Sience*, *34*(3), 521–532. doi: 10.1111/j.1551-6709.2009.01078.x

Szymanik, J., & Zajenkowski, M. (2011). Contribution of working memory in parity and proportional judgments. *Belgian Journal of Linguistics*, *25*(1), 176–194. doi: doi:10.1075/bjl.25.08szy

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*. Retrieved from http://www.illc.uva.nl/AC/AC2013/Proceedings

Thorne, C. (2012). Studying the distribution of fragments of English using deep semantic annotation. In *Proceedings of the Eighth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation* (pp. 66–71). Retrieved from http://sigsem.uvt.nl/isa8/ISA-8_proceedings-2.pdf

Tian, Y. (2014). *Negation Processing: A Dynamic Pragmatic Account* (Doctoral dissertation, University College London). Retrieved from http://discovery.ucl.ac.uk/1434202/

Tian, Y., Breheny, R., & Ferguson, H. J. (2010). Why we simulate negated information: A dynamic pragmatic account. *The Quarterly Journal of Experimental Psychology*, *63*(12), 2305–2312. doi: 10.1080/17470218.2010.525712

Tomaszewicz, B. (2013). Linguistic and visual cognition: Verifying proportional and superlative Most in Bulgarian and Polish. *Journal of Logic, Language and Information*, *22*(3), 335–356. doi: 10.1007/s10849-013-9176-6

Troiani, V., Clark, R., & Grossman, M. (2011). Impaired verbal comprehension of quantifiers in corticobasal syndrome. *Neuropsychology*, *25*(2), 159–165. doi: 10.1037/a0021448

Troiani, V., Peelle, J. E., Clark, R., & Grossman, M. (2009). Is it logical to count on quantifiers? Dissociable neural networks underlying numerical and logical quantifiers. *Neuropsychologia*, *47*(1), 104–111. doi: 10.1016/j.neuropsychologia.2008.08.015

Troiani, V., Peelle, J. E., McMillan, C., Clark, R., & Grossman, M.

(2009). Magnitude and parity as complementary attributes of quantifier statements. *Neuropsychologia*, *47*(12), 2684–2685. doi: 10.1016/j.neuropsychologia.2009.04.025

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. doi: 10.1016/j.jml.2015.03.010

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. doi: 10.1016/j.jml.2010.03.008

van Benthem, J. (1986). *Essays in Logical Semantics*. Dordrecht: Reidel.

van Dalen, D. (1994). *Logic and Structure*. Berlin: Springer. doi: 10.1007/978-3-662-02962-6

van Lambalgen, M., & Hamm, F. (2005). *The Proper Treatment of Events*. Malden, MA: Blackwell.

van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science*, *32*(6), 939–984. doi: 10.1080/03640210801897856

van Rooij, I., Kwisthout, J., Blokpoel, M., Szymanik, J., Wareham, T., & Toni, I. (2011). Intentional communication: Computationally easy or difficult? *Frontiers in Human Neuroscience*, *5*. (article 52) doi: 10.3389/fnhum.2011.00052

Vázquez, C. A. (1981). Sentence processing: Evidence against the serial, independent stage assumption. *Journal of Psycholinguistic Research*, *10*(4), 363–374. doi: 10.1007/BF01067164

Veale, T., & Keane, M. (1997). The competence of sub-optimal theories of structure mapping on hard analogies. In *Proceedings of the 15th International Joint Conference on Artifical Intelligence* (pp. 232–237). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: a neural model. *Journal of Cognitive Neuroscience*, *16*(9), 1493–1504.

von der Malsburg, T., & Angele, B. (2015). False positives and other statistical errors in standard analyses of eye movements in reading. *ArXiv e-prints*. Retrieved from http://arxiv.org/abs/1504.06896

von Stechow, A. (1984). Comparing semantic theories of comparison. *Journal of Semantics*, *3*(1-2), 1–77. doi: 10.1093/jos/3.1-2.1

Wabersich, D. (2014). Rwiener: Wiener process distribution functions [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=RWiener (R package version 1.2-0)

Wareham, H. T. (1999). *Systematic Parameterized Complexity Analysis in Computational Phonology* (Doctoral dissertation, University of Victoria, Victoria, B.C.). Retrieved from http://www.nlc-bnc.ca/obj/s4/f2/dsk2/ftp02/NQ37368.pdf

Weber, H. J., & Arfken, G. B. (2003). *Essential Mathematical Methods for Physicists*. Amsterdam: Academic Press.

Wellwood, A. (2015). On the semantics of comparison across categories. *Linguistics and Philosophy*, *38*(1), 67–101. doi: 10.1007/s10988-015-9165-0

Westerlund, M., Kastner, I., Al Kaabi, M., & Pylkkänen, L. (2015). The LATL as locus of composition: MEG evidence from English and Arabic. *Brain and Language*, *141*, 124–134. doi: 10.1016/j.bandl.2014.12.003

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85. doi: 10.1016/0001-6918(77)90012-9

Zajenkowski, M., Styła, R., & Szymanik, J. (2011). A computational approach to quantifiers as an explanation for some language impairments in schizophrenia. *Journal of Communication Disorders*, *44*(6), 595–600. doi: 10.1016/j.jcomdis.2011.07.005

Zajenkowski, M., & Szymanik, J. (2013). MOST intelligent people are accurate and SOME fast people are intelligent. *Intelligence*, *41*(5), 456–466. doi: 10.1016/j.intell.2013.06.020

Zajenkowski, M., Szymanik, J., & Garraffa, M. (2014). Working memory mechanism in proportional quantifier verification. *Journal of Psycholinguistic Research*, *43*(6), 839–853. doi: 10.1007/s10936-013-9281-3

Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation*, *10*(2), 403–430. doi: 10.1162/089976698300017818