Doctoral theses at NTNU, 2023:204

Heming Strømholt Bremnes

# Quantifiers and Complexity

## The distinct neural signatures of verification algorithms for natural language quantifiers.

Doctoral thesis

**NTNU**
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Humanities
Department of Language and Literature

NTNU
Norwegian University of
Science and Technology

Heming Strømholt Bremnes

# Quantifiers and Complexity

## The distinct neural signatures of verification algorithms for natural language quantifiers.

Thesis for the Degree of Philosophiae Doctor

Trondheim, June 2023

Norwegian University of Science and Technology
Faculty of Humanities
Department of Language and Literature

NTNU
Norwegian University of
Science and Technology

**NTNU**
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Humanities
Department of Language and Literature

ISBN 978-82-326-7110-6 (printed ver.)
ISBN 978-82-326-7109-0 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:204

Printed by NTNU Grafisk senter

Til Aleksander Hammer
(1987–2019)

# Abstract

This thesis concerns the impact of computational complexity on the verification of natural language quantifiers. In particular, it deals with the neural consequences of so-called minimal complexity: the simplest possible algorithm to compute a function. It can be shown mathematically that quantifiers in natural language divide into classes depending on the minimal complexity of determining the truth value of a quantified sentence with that quantifier: *Aristotelian* quantifiers include 'all', 'some', 'none' etc.; *numerical* quantifiers refer to numerals – like 'three', 'four', 'five' etc. – and modifications thereof; *parity* quantifiers concern the parity of a set, i.e. 'an even/odd number of'; and *proportional* quantifiers denote proportions, e.g. 'most', 'less than half', 'a third' etc.

Importantly, the first three classes can be verified by simple *finite state automata* (FSAs), whereas proportional quantifiers require the additional memory component associated with *pushdown automata* (PDAs). Since the formal proofs delineate a lower bound on the complexity of the verification algorithm – there is no strategy that can simplify the nature of the task – this leads to the prediction that proportional quantifiers, but not the other classes, should recruit memory systems during verification, also in human subjects.

In three EEG experiments with a picture-sentence verification paradigm, the cognitive reality of this distinction in complexity is investigated. These experiments demonstrate that the computational complexity of the verification algorithm for natural language quantifiers is reflected in distinct neural responses: proportional quantifiers led to specific effects in the event-related potential (ERP) compared to Aristotelian and numerical quantifiers, at different positions in a sentence in which verification occurs; moreover, these distinct effects for proportional quantifiers were modulated by overall memory load in a task that, in addition to verification, required participants to temporarily store and recall strings of digits.

This compelling evidence suggests that human language processing is subjected to the same constraints as those applicable to abstract machines. On the basis of these results, the thesis goes on to explore open questions at the intersection of computer science and psycholinguistics, and asks how and whether formal proofs about the complexity of specific computational problems can inform us about which class of algorithms is plausibly implemented by the brain. More generally, the thesis should be viewed as a proof of concept for a growing literature advocating algorithmic and complexity theoretic analyses in the construction of psychological and psycholinguistic theories.

# List of papers

1. Bremnes, H. S., Szymanik, J., and Baggio, G. (2022). Computational complexity explains neural differences in quantifier verification. *Cognition.* 223. doi: 10.1016/j.cognition.2022.105013

2. Bremnes, H. S., Szymanik, J., and Baggio, G. (Under revision). The interplay of computational complexity and memory load during quantifier verification.

3. Bremnes, H. S. (Submitted). Neural Algorithms of Natural Language Quantification: A review of the experimental literature

## Note on papers 1 and 2

Papers 1 and 2 were co-authored with Jakub Szymanik and Giosuè Baggio. In both cases, Szymanik and Baggio contributed to the design of the study, the interpretation of the results, and the preparation of the original draft. Additionally, Baggio oversaw the data analysis.

# Acknowledgements

In the late summer of 2016, I met with Giosuè Baggio, who was to become my primary supervisor, to discuss a potential PhD application due only three weeks later. I told him that I envisioned a project involving set theory and neuroscience, upon which he promptly gave me a fresh off the press book, written by Jakub Szymanik, who was to become my secondary supervisor. On the newfangled pages of this book lay the origins of what was to consume my academic life for the years to follow. While nothing came of this specific PhD application – hardly unexpected, given the time frame – looking back, this meeting seems like the wheels of a finely tuned clockwork coming together. The book contained a project waiting to happen, this project was exactly the sort of project I had envisioned for my PhD, and the two people who came to supervise me, were exactly the people I needed to be able to carry out the project.

I would therefore like to begin by sincerely thanking both my supervisors, who with their deliberative supervision style, in which topics are discussed in a principled and reasoned manner, have made me feel like I am an equal partner and an independent scientist.

Secondly, I have to thank everyone who participated in my experiments and endured the tedious monotony of differently colored circles and triangles accompanied by trivial statements. Relatedly, I wish to thank everyone who helped me recruit these participants, either by sharing the ads or by nagging family and friends. There would be no thesis without either of you.

Sincere gratitude is also due to Aniello De Santo for reading and commenting on an earlier draft of this thesis. Both his pertinent corrections and his exciting ideas for future research scrambled in the margins, have made the last few months more enjoyable than they otherwise would have been.

I would also like to extend my appreciation to my colleagues at the Department of Language and Literature for usually having their doors open and always being willing to answer my random academic – and other – queries. The same is true for the administrative staff both within the department and at the Faculty of Humanities, who have promptly resolved all non-academic problems and created a welcoming atmosphere where everyone feels cared for. However, those who have made me feel the most at home in the department, is the the PhD and young researcher community. I would therefore like to express a heartfelt thank you to everyone who have labored alongside me at various stages of my time as a doctoral research fellow; you have made me want to go to work, caused me to take too long breaks, and balanced out the meticulous scientific endeavour with laughter and conversations with friends.

Lastly, I would like to thank my family for their support and, amazingly, continued interest in my research. In particular, my deepest gratitude goes to my wife, Tina, and my two boys, Olav and Åsmund, who have forced me to optimize the algorithm for completing the PhD, and shown me that there is more to meaning than verification.

Trondheim, December 2022
Heming Strømholt Bremnes

# Table of Contents

# Chapter 1

# Introduction

Natural language quantifiers are linguistic expressions that denote quantities. The morphosyntactic realization of quantifiers exhibit substantial differences, spanning from monomorphic determiners such as ‘all’ and ‘five’, via modified quantity nouns like ‘at most two thirds’ and ‘an even number of’, to multiple conjoined phrases: ‘at least two but no more than five’ or ‘between five and seven or more than ten’. Despite the apparent diversity of these linguistic expressions, their semantic contribution is importantly similar. On the regular mathematical definition of quantifiers in formal semantics (to be discussed in 1.1), quantifiers denote relations between cardinalities of sets. Interestingly, natural language quantifiers constitute a small subset of the logically possible quantifiers *qua* cardinality relations (Barwise & Cooper, 1981; Keenan & Stavi, 1986), and quantifier expressions have been shown to be remarkably invariant cross-linguistically, both in terms of meaning and form (Bach, Jelinek, Kratzer, & Partee, 1995; Keenan & Paperno, 2017; Matthewson, 2001). Furthermore, the logical properties that characterize the subset of natural language quantifiers (in the universe of possible quantifiers) seem to delineate learning biases for quantitative tasks in non-human primates (Chemla, Dautriche, Buccola, & Fagot, 2019), as well as quantifier learning more generally (Carcassi, Steinert-Threlkeld, & Szymanik, 2021; Hunter & Lidz, 2013; Steinert-Threlkeld & Szymanik, 2019; van de Pol, Steinert-Threlkeld, & Szymanik, 2019). These facts suggest that studying the neurobiology of quantifiers can reveal fundamental truths about the human capacity for language in general, the place of language in the human mind, and the evolution of the human cognitive architecture from the primate brain.

The study of brain activity is, however, associated with a number of conceptual problems. The detection of increased metabolism in areas of the cortex through functional magnetic resonance imaging (fMRI) or the observation of correlations between an external stimulus and patterns of electromagnetic activity measured at the scalp through electroenchephalography (EEG) or magnetoencephalography (MEG) does not in and of itself answer such constitutional questions. Since explanations are answers to 'why'-questions (Garfinkel, 1981; Lipton, 1991, 2004; van Fraassen, 1977, 1980), a theoretical account of the cognitive resources required to perform a task – e.g., comprehending quantified sentences – is necessary for neural data to be explanatory. In other words, the theory explains the activation patterns observed in neurobiological experiments and thereby the cognitive capacity one is researching.

In *Vision*, Marr (1982) famously argued that in computational cognitive science, three levels of analysis are necessary to explain information processing systems. *The computational level* is specifying a procedure as a function, i.e. as an input-output mapping. *The algorithmic level* describes the stepwise computation of this function, i.e. the procedure for transforming inputs into outputs. Finally, *the implementational level* details how this algorithm is instantiated in the biological substrate, i.e. how the brain performs the algorithm that allows it to compute a function. While all these levels are important, the true explanatory power of this framework comes from the inferences that can be drawn between the levels. For example, if a function is not tractable – i.e. computable in realistic time – this cannot be the function that the brain is computing (van Rooij, 2008; van Rooij, Blokpoel, Kwisthout, & Wareham, 2019). Likewise, if a person computes a function in a certain amount of time, an algorithm that cannot be computed as quickly given the processing power available in the brain, cannot be the algorithm that the brain is instantiating (Carruthers, Stege, & Masson, 2018).

Being constrained both by the function that needs to be computed and by the limitations imposed by the physical medium of the brain, the algorithmic level can be viewed as the mediator between the computational and implementational levels (Baggio, Stenning, & van Lambalgen, 2016; Baggio, van Lambalgen, & Hagoort, 2015; Embick & Poeppel, 2015; S. Lewis & Phillips, 2015). Still the algorithmic level has not received sufficient attention in semantics (Baggio, 2018, 2020). One reason for this lack of consideration might be the difficulty in formalizing linguistic meanings, seeing as formalization is necessary to state a problem in computational terms. Coincidentally, this is not the case for quantifier expressions.

As well as being important in natural language and cognition, quantifiers lie at the heart of logic and mathematics, and this thesis seeks to capitalize on this parallelism to explain how quantification is realized in the brain. More precisely, its aim is to apply a synthesis of the findings on quantifiers in logic, computability theory, and linguistics, to the neural processing of quantifiers. In particular, it can be shown mathematically that quantifiers can be divided into four classes depending on the computational resources required to verify them (M. Mostowski, 1998; Szymanik, 2016; van Benthem, 1986c). The reason for focusing on verification is twofold. Firstly, verification is a well-defined function, where sentences are mapped to truth values, given a context. Secondly, truth values have an elevated status within formal semantics, since, conventionally, the extension of a declarative sentence is its truth value. As a consequence of this prominence, the results about quantifier verification potentially have far reaching consequences for the meaning and processing of quantifiers. For example, *procedural semantics* (Moschovakis, 2006; Muskens, 2005; Pietroski, Lidz, Hunter, & Halberda, 2009; Suppes, 1982; Szymanik, 2016; Tichý, 1969; van Benthem, 1986b; van Lambalgen & Hamm, 2005) contends that the meaning of an expression is a set of algorithms computing its extension, and consequently that the meaning of quantified sentences is a set of verification algorithms. However, the work presented in this thesis, remains agnostic on this philosophical position about meaning.

It is important to note that the formal proofs about quantifier verification do not implicate specific algorithms for sentence processing. Rather, they identify properties of classes of algorithms and delineate a lower bound for the verification complexity of different quantifiers (so-called *minimal complexity*), and to an extent, which resources need to be recruited in order to verify certain classes of quantifiers, but not others. As will be discussed in more detail in 1.1 below, as well as in chapter 2, the identification of these properties relies on certain assumptions. In particular, the formal proofs assume that verification is sequential and exact, i.e. that people enumerate all the objects in the domain of the quantifier individually. For that reason, these idealized algorithms of quantifier verification are arguably not a realistic model of human performance on such tasks, at least not in all contexts.

While the formal proofs may not be applicable to human verification in all situations, this abstraction can still be considered informative: Since there are many different algorithms people might employ to verify a certain expression, what is important is the characteristic properties of these algorithms. As Niyogi (2006, p. 39) writes: “for mathematical models the assumptions are more questionable but the conclusions are more reliable – for computational models, the assumptions are more believable but the conclusions more suspect.” For example, while the quest for quantifier specific algorithms explored in a parallel literature (e.g. Hackl, 2009; Hunter, Lidz, Odic, & Wellwood, 2017; Knowlton et al., 2021; Lidz, Pietroski, Halberda, & Hunter, 2011; Pietroski et al., 2009; Pietroski, Lidz, Hunter, Odic, & Halberda, 2011; Talmina, Kochari, & Szymanik, 2017; Tomaszewicz, 2011) is interesting and informative in its own right, it fails to appreciate commonalities between such algorithms. Similarly, plausible models of human reasoning with quantifiers (e.g Khemlani & Johnson-Laird, 2022; Tessler, Tenenbaum, & Goodman, 2022) might inform us about the limitations of human reasoners, but the merit of such models depend on their ability to mirror behavioral and/or neural data, which in turn can be explained by properties described in mathematical proofs. This is to say that all these ventures have merit, and contribute to our understanding of quantifier processing in distinct ways.

Bearing that in mind, I will now go on to derive the different classes of natural language quantifiers informally. On the basis of this derivation, a few natural research questions will crystallize, and it is the aim of this thesis to begin to answer these.

## 1.1 Deriving Quantifier Classes

The semantics of natural language quantifiers, as hinted at initially, can be mathematically represented as relations between cardinalities of sets. While I will reserve the formal details of such representations to chapter 2, it is helpful to give a few illustrative examples in a colloquial manner. In order for a quantified sentence like 'All men are mortal' to be true, the set of men needs to be a subset of the set of mortal things. Equivalently, the cardinality of the set of men, i.e. the number of men, is equal to the cardinality of the set of men who are mortal. One can easily construct such examples for *numerical* or *parity* quantifiers as well, where sentences like 'three men are mortal' or 'an odd number of men are mortal' are true just in case the cardinality of the intersection of men and mortals equals 3 and the number of mortal men is an odd number, respectively. The relation denoted by *proportional* quantifiers, however, is slightly different. Consider a sentence like 'most men are mortal', despite its questionable implicature. As the name suggests, proportional quantifiers denote proportions. Specifically, they describe relations between the things that have a property and the things that do not, in this case men who are mortal and men who are not mortal. The sentence is true just in case the mortal men outnumber the immortal men.

Johan van Benthem (1986c) examined some interesting computational properties of natural language quantifiers *qua* relations between cardinalities of sets. If we represent a model of the relevant parts of the universe of discourse – for most quantifiers the set of A, for a quantified sentence Q AB – as a string of binary, where 1s represent the As that are B and 0s represent the As that are not, quantifiers provably fall into four distinct classes depending on the complexity of the algorithm required to recognize the string:

**Aristotelian**: 'All', 'some', 'not all', 'no'

**Numerical**: 'three', 'four', 'five',...

**Parity**: 'An even/odd number of'

**Proportional**: 'Most', 'More/less than half', 'a third', ...

Since quantifiers denote “families of sets” (Barwise & Cooper, 1981), i.e. the set of sets that make the quantifier expression true, a quantifier can also be said to denote a set of strings that represents these sets. The computational problem thus becomes determining whether a string belongs to the set of strings denoted by a specific quantifier. This is a foundational task in theoretical computer science, where one builds so-called *automata* – essentially an algorithm – that can solve this task (Hopcroft & Ullman, 1979), and van Benthem (1986b) thus aptly named his approach to quantifier meanings ‘Semantic Automata’.

Setting aside the abstract representations for a moment, we can illustrate these algorithms more intuitively with the circles and triangles that will become familiar to the reader throughout this thesis. All algorithms rely on going through the objects in the domain, e.g. the circles, sequentially, and determining for each object whether that object has the predicated property, e.g. being red. For a sentence like ‘all the circles are red’, the algorithm outputs true if, after scanning all the circles, it has not found a non-red circle. For ‘three triangles are yellow’, one ignores all the non-yellow triangles and counts the yellow until one reaches three, in wich case the sentence is true.[1] Both these classes of quantifiers can be verified using the simplest kind of automaton, an *acyclic finite state automaton* (acyclic FSA). Still, they are slightly different because the complexity of the algorithm for numerical quantifiers depends on the counting steps denoted by the numeral whereas the complexity of the Aristotelian quantifier algorithm is fixed. This is because for higher numbers – compare ‘three’ and ‘five hundred ninety seven’ – there is a more substantial counting procedure that requires more computational resources (Szymanik, 2016). Parity quantifiers can also be computed by an FSA, but in this case it needs to be cyclic (M. Mostowski, 1998). To exemplify, for a sentence like ‘an even number of circles is yellow’, the algorithm keeps track of whether the current number of yellow circles is odd or even. This obviously changes every time it sees a yellow circle. If the algorithm is in the even state, i.e. it has seen an even number of yellow circles, when it has inspected all the circles, the sentence is true.

By contrast, proportional quantifiers provably cannot be computed by an FSA, but requires the additional computational resources of a *pushdown automaton* (PDA) that has a memory stack (van Benthem, 1986c). This is because it is not enough to keep track of only the last object one has observed when verifying a sentence with a proportional quantifier. Consider the sentence ‘Less than half of the triangles are red’. In order to verify this sentence, the algorithm must keep track of the ratio of red to non-red triangles. It therefore stores the current ratio in memory and updates

[1] Numerical quantifiers famously have an *at least* and an *exact* reading (e.g. Levinson, 1983), and this algorithm gives you the *at least* reading. Since the complexity of the algorithm does not change depending on the reading, I will not discuss it here, but see chapter 2.2.2 for details.

it for every triangle it sees. If when all the triangles have been scanned, the non-red circles outnumber the red circles, the sentence is true.

So while there are four distinct verification classes of natural language quantifiers, the most prominent distinction is between proportional and non-proportional quantifiers. The minimally complex algorithm for proportional quantifiers is strictly more complex than the minimally complex algorithm for any class of non-proportional quantifiers. From the perspective of cognitive science, one interesting aspect of this complexity is that it is related to memory. At a minimum, the prediction of the semantic automata theory of quantifier verification is therefore that proportional quantifiers necessitate the recruitment of memory systems, whereas non-proportional quantifiers do not.

## 1.2 Research Questions

As mentioned in the introduction, the aim of this project is to apply formal results to further our understanding of the processing of natural language quantifiers. There is preliminary evidence to suggest that the minimal complexity of quantifier verification algorithms has real psychological and neural effects (De Santo, Rawski, Yazdani, & Drury, 2019; McMillan, Clark, Moore, Devita, & Grossman, 2005; McMillan, Clark, Moore, & Grossman, 2006; Morgan et al., 2011; Olm, McMillan, Spotorno, Clark, & Grossman, 2014; Szymanik, Meijering, & Verbrugge, 2013; Szymanik & Zajenkowski, 2010a, 2010b, 2011; Troiani, Peelle, McMillan, Clark, & Grossman, 2009a; Zajenkowski, Styła, & Szymanik, 2011; Zajenkowski & Szymanik, 2013; Zajenkowski, Szymanik, & Garraffa, 2014). However, the neural work is either small scale or has relied on spurious divisions between quantifier classes. It is therefore paramount to discover whether the predictions borne out in behavioral data has neural counterparts. Furthermore, the impact of verification complexity on the various stages of sentence processing has mostly not been examined at all. Since semantic theory dictates that truth values are important in sentence processing, manipulating verification complexity might provide a window into how they are important. The following research questions can be formulated to this end:

(1) Do the differences in the computational complexity of verification algorithms for proportional and non-proportional quantifiers manifest in distinct brain responses?

(2) If so, are these brain responses related to memory, as predicted by the automata theory?

(3) At what point(s) during sentence processing do such differences occur?

(4) What model of sentence processing best explains such patterns?

If (1) can be answered in the affirmative, it will demonstrate the predictive power of hypotheses derived from analyzing cognitive tasks in computational terms. In particular, it would suggest that we can understand crucial aspects of the processing of quantifiers on the basis of formal results alone. Research question (2), even if framed in binary terms, is slightly more nuanced. Mathematical theories are independent of the linking hypotheses that tie them to cognitive systems (van Rooij & Baggio, 2021), and there is therefore more than one way in which the brain responses can be related to memory. At most, the theory predicts that *a* memory component should distinguish proportional from non-proportional quantifiers, and the implementational details of this abstract notion of memory needs to be inferred from the psychology of verification, and ultimately from the extant empirical results.

Considering that human sentence processing happens rapidly, at the scale of milliseconds, (3) places certain restrictions on the methodological choices in that the measurements of neural activity needs to have the correct temporal resolution. Since fMRI relies on the hemodynamic response and therefore the speed of bloodflow in the homeostatic circulatory system, it does not have the capability to detect differences between various stages of sentence processing. Consequently, it is necessary to utilize electrophysiological methods such as EEG in order to answer the research questions. The answer to (3) has consequences for (4) as well. Depending on whether differences in the verification procedure are observed early or late in the sentence, it is possible to discern whether participants are actively building true sentences incrementally, predicting the upcoming verbal material, or whether they wait and only attempt to verify the sentence once a complete proposition is available.

## 1.3 Outline of the Thesis

An elaboration on the derivation of the quantifier classes, with the formal underpinnings of generalized quantifier theory and its operationalization into algorithms is found in chapter 2. Chapter 3 deals with methodological considerations. This involves both discussing how cognitive neuroscience ought to be informed by formal theory, and considering the more concrete implementation of the research questions into a valid experimental paradigm. A summary of the papers is provided in chapter 4, before the results are synthesized in order to answer the research questions in chapter 5. The thesis ends with chapter 6, where I contemplate the contribution of this thesis to the larger research community and suggest directions for future research.

# Chapter 2

# Quantifiers and Their Associated Computational Profiles

In order to demonstrate that quantifiers fall into different classes depending on their computational complexity, some formal prerequisites are required. This chapter is devoted to providing these prerequisites. Firstly, the semantics of quantifiers is specified in section 2.1. Secondly, how to turn these semantics into algorithms is described in 2.2, before, finally, the different quantifier classes are derived from such algorithms in 2.3.

## 2.1 Generalized Quantifier Theory

While quantification is as old as logic itself, both the Aristotelian and Fregean logics were primarily made for the quantifiers used in syllogistic reasoning, i.e. $\forall$ and $\exists$. However, in two seminal papers, A. Mostowski (1957) and Lindstrøm (1966) provided the mathematical framework that laid the groundwork for the later introduction of formal accounts of, potentially, all natural language quantifiers in semantics (e.g. Barwise & Cooper, 1981; Gärdenfors, 1987; Keenan & Stavi, 1986; van Benthem & ter Meulen, 1985; Westerståhl, 1985). The groundbreaking idea was to define the semantics of quantifiers in terms of relations between cardinalities of sets. I will not provide a full first-order logic with generalized quantifiers, nor a compositional semantics for how these meanings are derived from syntax, as this is beyond the prequisites for the results in this thesis; see however Peters and Westerståhl (2006) and Heim and Kratzer (1998), respectively. I will begin by defining what generalized quantifiers are in 2.1.1, before demonstrating how this can be applied to natural language quantifiers in 2.1.2. Finally, I will describe some relevant logical properties of natural language quantifiers in 2.1.3.

### 2.1.1 Formal definitions

In the following, I assume familiarity with set notation and basic first order logical connectives. On a further notational note, a *vocabulary* is a finite set V of symbols denoting predicates of various arities. Let $\tau = \{R_1, ..., R_k\}$ be a vocabulary, where $n_i$ is the arity of $R_i$, for each i. A model of $\tau$ is a structure given by $\mathbb{M} = (M, R_1, ..., R_k)$ where M is the universe of model $\mathbb{M}$ and $R_i \subseteq M$ is an $n_i$-ary relation over M, for $1 \leq i \leq k$.

**Definition 2.1.1.** Let $t = (n_1, ..., n_k)$ be a $k$-tuple of positive integers. A *generalized quantifier* of type $t$ is a class Q of models of a vocabulary $\tau_t = \{R_1, ..., R_k\}$, such that $R_i$ is $n_i$-ary for $1 \leq i \leq k$. Q is isomorphism closed, such that if $\mathbb{M}$ and $\mathbb{M}'$ are isomorphic, then

$$(\mathbb{M} \in Q \iff \mathbb{M}' \in Q).$$

**Definition 2.1.2.** If on each universe M, Q is a relation between subsets of M, i.e. if it is of type $\langle 1, ..., 1\rangle$, Q is *monadic*. Otherwise it is *polyadic*.

In more colloquial terms, a generalized quantifier corresponds to a class of models, consisting of relations between relations of a universe M in a model $\mathbb{M}$.[1] What class of models is denoted by the quantifier will depend on the specific quantifier, and which relations are in the universe, will, for our purposes, be given by the surrounding linguistic context. For example, the denotation of noun phrase quantifiers, such as ‘someone’ or ‘everything’, are type $\langle 1\rangle$ quantifiers that are devoid of linguistic modification, and therefore denote a property, a unary predicate, on the universe, M: $Q_{someone} = \{A \subseteq M : A \neq \emptyset\}$; $Q_{everything} = \{M\}$. However, the most basic type of quantifiers in natural language (Peters & Westerståhl, 2006, p. 12) are type $\langle 1, 1\rangle$ quantifiers, linguistically expressed as determiners, like ‘all’, ‘some’, ‘three’ and ‘most’. These are by far the most common, and most studied, quantifiers, and are attested in every language (Keenan & Paperno, 2017). As might be inferred from the typing, the denotation of such quantifiers express relations between sets of individuals. The specific relations between sets instantiated by such quantifiers will be dealt with in more detail in 2.1.2 below.

A distinction worth mentioning before eventually setting it aside, is the distinction made by Partee (1995) between D-quantifiers and A-quantifiers. All quantifiers discussed so far have been D-quantifiers, which are quantifiers that quantify over objects and modify noun phrases. By contrast, A-quantifiers quantify over events and generally function as adverbials. In English, such quantifiers include lexicalized adverbs like ‘always’ and ‘seldom’, but A-quantifiers are also commonly derived

[1] Defining generalized quantifiers in terms of classes of models is equivalent to the definition of a generalized quantifier as a functional relation between relations that might be more familiar to the reader (Szymanik, 2016, Corollary 3.1).

from D-quantifiers, e.g. ‘five/most times’. Importantly, the converse does not hold cross-linguistically: there are no known examples of D-quantifiers being derived from A-quantifiers, and while all languages have simple A-quantifiers, A-quantifiers are generally more morpho-syntactically complex than D-quantifiers (Gil, 1993; Keenan & Paperno, 2017). Relatedly, aside from the entities they denote and their typical syntactic function, A-quantifiers and D-quantifiers are mathematically, and consequently computationally, the same. Additionally, and more importantly in the context of the experimental nature of this thesis, A-quantifiers seem to be associated with the same kind of electrophysiological effects (Augurzky, Hohaus, & Ulrich, 2020). In what follows I will therefore only be concerned with D-quantifiers.

Since the quantifier types discussed so far only deal with relations between unary predicates, they are all monadic quantifiers. A lot of time has been devoted to studying the possibilities for polyadic quantification in natural language (Keenan, 1987, 1992; Keenan & Moss, 1985; Moltmann, 1992, 1995; Nam, 2005; van Benthem, 1989) and the possibility to reduce them to monadic quantifiers. The main finding from these works is that there are quantifiers that have proven irreducible. Among these are reciprocals, that denote relations between a unary predicate and a relation, and are consequently of type $\langle 1, 2\rangle$:

1. Every student admires himself. (Keenan, 1987)
2. The ten students criticized each other. (Keenan, 1987)
3. The candidates criticized each other and each other’s wives (Keenan, 1992)

Same and different comparisons, branching quantifiers, and exception anaphora, take two sets and express a relation between them, and are of type $\langle 1, 1, 2\rangle$:

4. Most of the students answered the same number of questions on the exam.
5. Each teacher assigned a different number of problems to the same student.
6. Most philosophers and most linguists agree with each other about branching quantification. (Barwise, 1979)
7. Every man danced with every woman except John with Mary. (Moltmann, 1995)

However, this falls beyond the scope of the present thesis, in particular since there are open problems concerning how these quantifiers relate to semantic automata (Steinert-Threlkeld & Icard III, 2013; Szymanik, 2016; Szymanik, Steinert-Threlkeld, Zajenkowski, & Icard III, 2013).

### 2.1.2 A Semantics for Certain Quantifiers

As promised, I will now return to the semantics for type $\langle 1, 1\rangle$ determiners. This section relies heavily on the groundwork in Barwise and Cooper (1981), as this is where the canonical semantics for most natural language quantifiers are found. We have already seen the denotations for some type $\langle 1\rangle$ Aristotelian quantifiers, so let us first extend universal and existential quantification to relations between two sets A and B, with the addition of the other Aristotelian quantifiers 'no' and 'not all':

$$Q_{all}A, B \iff A \subseteq B$$

$$Q_{some}A, B \iff A \cap B \neq \emptyset$$

$$Q_{no}A, B \iff A \cap B = \emptyset$$

$$Q_{not\,all}A, B \iff A \nsubseteq B$$

We see that the meaning of these quantifiers can be captured exlusively by reference to subset relations and the empty set. For example, 'all' and 'not all' when used in sentences like '(not) all A are B' specifies whether the set of As is or is not a subset of B. 'Some' and 'no' in comparable sentences states that the intersection of A and B is or is not the empty set.

However, it turns out that these relations are the exceptions rather than the rule. For most natural language quantifiers, the quantifier specifies a relation between the cardinalities of A and B:

$$Q_{exactly\,three}A, B \iff |A \cap B| = 3$$

$$Q_{at\,least\,five}A, B \iff |A \cap B| \geq 5$$

$$Q_{an\,even\,number\,of}A, B \iff 2 \mid |A \cap B|$$

$$Q_{an\,odd\,number\,of}A, B \iff 2 \nmid |A \cap B|$$

$$Q_{most}A, B \iff |A \cap B| \geq |A - B|$$

$$Q_{less\,than\,half}A, B \iff |A \cap B| \leq \frac{|A|}{2}$$

$$Q_{a\,third\,of}A, B \iff |A \cap B| = \frac{|A|}{3}$$

As is evident from the above denotations, there are various relations that can be said to hold of the cardinality of the intersection between A and B. One can readily refer to the exact cardinality by using a *numerical* quantifier, either as a threshold ('at least') or precisely ('exactly'). Bare numerals are famously ambiguous between the two readings (e.g. Horn, 1972; Levinson, 1983), and I have therefore written denotations with a modifier for clarity. Many languages can also refer to the parity of a cardinality by using 'an even/odd number of' as a quantifier. In such cases, there is a relation between the number 2 and the cardinality of the intersection of A and B: either 2 divides or does not divide the cardinality.[2] The remaining denotations above are all examples of *proportional* quantifiers. These all denote relations between the cardinality of the intersection of A and B and some other cardinality. For example, 'most As are B' is true if there are more As that are B than As that are not B, i.e. the cardinality of the intersection is greater than the cardinality of the difference of A and B. When the proportion is specified, as in 'less than half' and 'a third of', the cardinality of the intersection of A and B is said to be greater/less than or equal to some proportion of the cardinality of only A, in this case half or a third.

It is pertinent to mention that the denotation for 'most' is subject to debate. Importantly, Hackl (2009) argued convincingly that there is a *superlative* and a *proportional* reading for 'most', where the proportional reading is the one given above, which is equivalent to the corresponding denotation for 'more than half'. However, on a superlative reading, 'most' functions as a regular superlative, comparing the denotation of the quantified noun-phrase to a *comparison class* (Farkas & Kiss, 2000; Heim, 1999; Sharvit & Stateva, 2002; Tomaszewicz-Özakın, 2020). This difference also manifests in separate lexicalized quantifiers in certain languages (Tomaszewicz, 2011). Importantly, Hackl (2009) points out that the absolute reading of superlatives corresponds to the proportional quantifier meaning, and that 'most' is therefore better analyzed as an adjective than a quantifier. As further evidence for this, Hackl argues that the negative polarity counterpart to 'most', 'fewest', can only have the superlative reading, since the absolute reading - *smallest subset* - is infelicitous because there is no smallest subset. Nevertheless, it is possible to give generalized quantifier semantics for both these readings, and while 'fewest' and 'less than half' are clearly not synonymous, in universes where all As are either B or not B, these two quantifiers have the same denotation, and 'fewest' can therefore be seen as the polar opposite to 'most' in such universes.[3]

Hackl's (2009) work also set in motion a debate about differences between 'most' and 'more (than half)' (e.g. Carcassi & Szymanik, 2021; Denić & Szymanik, 2022; Knowlton et al., 2021; Lidz et al., 2011; Ramotowska, Steinert-Threlkeld, van Maanen, & Szymanik, 2020; Talmina et al., 2017). In short, 'most' appears to connote

[2] Note that parity quantifiers are an instance of the wider class of *divisibility* quantifiers, i.e. quantifiers denoting cardinalities divisible by $n$. However, such quantifiers mostly belong in mathematical logic (see e.g. M. Mostowski, 1991, for a technical definition), and since this thesis is concerned with natural language, discussing parity quantifiers will suffice.

[3] I will not discuss the felicity of utterances with 'fewest' in English here, since my experiments were conducted in Norwegian, where such sentences are less marked. For an overview of quantity superlatives in Germanic languages, the reader is referred to Coppock (2019).

a larger proportion than 'more than half', while at the same time being associated with a difference in verification strategy, that is better represented by a denotation along these lines:

$$Q_{most}A, B \iff |A \cap B| \geq |A| - |A \cap B|$$

However, as these denotations are truth functionally equivalent, they do not have a bearing on the complexity of the verification procedures described below, and I will not pursue this any further.

Since most quantifiers require semantics in terms of relations between the cardinalities of sets, it might, for symmetry, be conducive to revisit the semantics for Aristotelian quantifiers, to show that it is possible to give semantics for the these quantifiers in terms of cardinality as well:

$$Q_{all}A, B \iff |A \cap B| = |A|$$

$$Q_{some}A, B \iff |A \cap B| \neq 0$$

$$Q_{no}A, B \iff |A \cap B| = 0$$

$$Q_{not\,all}A, B \iff |A \cap B| \neq |A|$$

As a final note on type $\langle 1, 1 \rangle$ determiners, I would briefly like to mention *value judgement quantifiers*, i.e. 'few', 'many', 'enough' etc. Such quantifiers are similar to numerical quantifiers in that they denote a cardinality greater than a certain value, call it $d$. The key difference is that $d$ is relative, and its value is determined by pragmatic or otherwise contextual factors (Rett, 2018). Furthermore, value judgement quantifiers can receive a *cardinal* – i.e. greater than a specific number – a *proportional* – greater than a certain proportion of A – or a *reverse proportional* – greater than a certain proportion of B – reading. I give the semantics for 'many' on all these three readings below:

$$Q^{cardinal}_{many}A, B \iff |A \cap B| > d$$

$$Q^{proportional}_{many}A, B \iff \frac{|A \cap B|}{|A|} > d$$

$$Q^{reverse\,proportional}_{many}A, B \iff \frac{|A \cap B|}{|B|} > d$$

On a cardinal reading of sentences like 'Many As are B', the number of As that are B is said to be larger than the number that could be expected. On a proportional reading, a larger proportion of As are B than what is expected, whereas on a reverse proportional reading, there is a bigger proportion of A that are B, compared to other things that are B, relative to some expected proportion. All these readings are attested in language, and they have some interesting properties that will be discussed briefly below. However, as these formalizations are subject to debate, and their meanining is determined in non-semantic ways, I will not discuss such quantifiers any further. For an overview, see Rett (2018), for discussions of formal properties, see Hackl (2000); Partee (1989); Romero (1998, 2015); or Westerståhl (1985), and for a discussion of contextually determined value judgments, see Cresswell (1976) and Rett (2015).

### 2.1.3 Properties of Natural Language Quantifiers

I will now present and define some properties of natural language quantifiers that are relevant when defining algorithms of quantifier verification. This means that there are further properties that are interesting for other purposes, but I will not discuss these in any detail. For a comprehensive overview, see Peters and Westerståhl (2006).

Firstly, it is easy to observe that quantifiers occur in complex noun phrases – e.g. 'most or all', 'between 5 and 10' – that modify the meaning of the quantifier. Such modification occurs in all languages (Keenan & Paperno, 2017), and it is therefore necessary to define Boolean combinations of quantifiers.

**Definition 2.1.3.** Let Q, Q' be generalized quantifies, both of type $\langle n_1, ..., n_k \rangle$. We define conjunction:

$$(Q \wedge Q')_M(R_1, ..., R_k) \iff Q_M(R_1, ..., R_k) \text{ and } Q'_M(R_1, ..., R_k)$$

Disjunction:

$$(Q \vee Q')_M(R_1, ..., R_k) \iff Q_M(R_1, ..., R_k) \text{ or } Q'_M(R_1, ..., R_k)$$

Outer negation:

$$(\neg Q)_M(R_1, ..., R_k) \iff \text{not } Q_M(R_1, ..., R_k)$$

Inner negation:

$$(Q\neg)_M(R_1, ..., R_k) \iff Q_M(R_1, ..., R_{k-1}, M - R_k)$$

Dual:

$$Q^d = \neg(Q\neg) = (\neg Q)\neg$$

It might also be helpful to define negation for type $\langle 1, 1 \rangle$ quantifiers, as this is what we are concerned with in what follows.

**Definition 2.1.4.** Let Q be a generalized quantifier of type $\langle 1, 1 \rangle$ and A, B $\subseteq$ M. We define outer and inner negation as below:

$$(\neg Q)_M(A,B) \iff \text{not } Q_M(A,B)$$

$$(Q\neg)_M(A,B) \iff Q_M(A, M - B)$$

This allows us to account for sentences such as:

8. Between 5 and 10 students attended the lecture.

9. Most or all students enjoyed formal semantics.

10. Not all students passed the exam.

11. Some students did not pass the exam.

It also allows us to see that the quantifiers in the following sentences are duals.

12. Not all students did not pay attention.

13. Some students payed attention.

More importantly in the present context, in particular in paper 3, is the fact that conjunction and negation of natural language quantifiers is used to define so-called *scalar implicatures*, typically associated with many quantifier expressions (Horn, 1972; Levinson, 1983). Such implicatures are defined as the inferred negation of a stronger meaning. For example, the quantifier 'some', in a sentence like $Q_{some}(A,B)$, frequently gives rise to the implicated meaning that not all the As are B. Consequently, in many instances $Q_{some}(A,B)$ in fact means $Q_{some}(A,B) \wedge \neg Q_{all}(A,B)$.

A property that is closely related to negation is *monotonicity*. A quantifier can be monotone on either of its arguments, and:

**Definition 2.1.5.** A quantifier $Q_M$ of type $(n_1, ..., n_k)$ is *monotone increasing in the i-th argument* if and only if:
If $Q_M[R_1, ..., R_k]$ and $R_i \subseteq R'_i \subseteq M^{n_i}$, then $Q_M[R_1, ..., R_{i-1}, R'_i, R_{i+1}, ..., R_k]$ where $1 \leq i \leq k$.

Conversely, $Q_M$ is *monotone decreasing in the i-th argument* if and only if:
If $Q_M[R_1, ..., R_k]$ and $R'_i \subseteq R_i \subseteq M^{n_i}$, then $Q_M[R_1, ..., R_{i-1}, R'_i, R_{i+1}, ..., R_k]$ where $1 \leq i \leq k$.

**Definition 2.1.6.** A quantifier is *monotone* if it is monotone increasing or decreasing in any of its arguments. Otherwise, it is nonmonotone.

Intuitively, this means that if a monotone quantifier is true of a set, it is also true of a superset, if it is monotone increasing, or of a subset, if it is monotone decreasing. For example, suppose a quantifier $Q(A,B)$ is monotone increasing on its second argument, and that $B \subseteq B'$. This means that if $Q(A,B)$ is true, then $Q(A,B')$ is also true. Or consider a quantifier $Q(A, B)$ that is monotone decreasing on its first argument, and $A' \subseteq A$. Then $Q(A, B)$ is true just in case $Q(A', B)$ is true. 14-17 give examples of quantifiers that are monotone increasing, monotone decreasing, nonmonotone, and of how monotonicity interacts with negation.

14. (a) At least 5 boys play football.
    (b) At least 5 boys play sports.

15. (a) At most 5 people came to the show.
    (b) At most 5 women came to the show.

16. (a) Exactly 5 boys play football.
    (b) Exactly 5 boys play sports.

17. (a) All the politicians are furious.
    (b) All the politicians are angry.
    (c) Not all the politicians are furious.
    (d) Not all the politicians are angry.

14 is monotone increasing on the second argument. Therefore (a) entails (b), since sports is a superset of football. 15 is monotone decreasing on its first argument, and since women are a subset of people, (a) entails (b). 16 is not monotone and so there is no entailment relation between the sentences. In 17 we see that negation flips the direction of the entailment. Without negation (a) entails (b), but with negation (c) does not entail (d). Rather, (d) entails (c).

While slightly peripheral to the present project, the fact that quantifiers can differ in their monotonicity is a confound that should be controlled for: There is considerable evidence that polarity – which is related to monotonicity in that only downward monotone quantifiers can be negative (see e.g. Fauconnier, 1975; Israel, 1996, 2001; Ladusaw, 1979) – affects processing (Clark & Chase, 1972, 1974; Deschamps, Agmon, Loewenstein, & Grodzinsky, 2015; Just & Carpenter, 1971; Nieuwland, 2016; Urbach, DeLong, & Kutas, 2015; Urbach & Kutas, 2010). More recent evidence also suggests that being monotone decreasing in itself is more strenuous for processing (Agmon, Loewenstein, & Grodzinsky, 2019; Geurts & van der Silk, 2005; Schlotterbeck, Ramotowska, van Maanen, & Szymanik, 2020). For a discussion of this evidence, see paper 3.

Recall from definition 2.1.1 above (cf. Lindstrøm, 1966; A. Mostowski, 1957), that generalized quantifiers are closed under isomorphism – or *topic neutral*. Informally, this means that quantifiers are only concerned with the cardinalities of sets of elements, and not the specific elements in the set or the order of these elements. However, there are potential counterexamples to isomorphism closure for natural language quantifiers. For example, while ‘three’ is isomorphism closed – if there is a one-to-one mapping between two models and $Q_{three}(A, B)$ is true of one of them, it is also true of the other – ‘first three’ or ‘every third’ are not: if two models have the same cardinality, but do not have the same ordering, ‘first three’ or ‘every third’ may be true in one and false in the other.

Less problematically, it was noted by van Benthem (1986a) that determiners in natural language, i.e. type $\langle 1, 1\rangle$ quantifiers, are *domain independent*, sometimes also called *extensional* (Ext). Formally, this can be expressed as follows:

**Definition 2.1.7.** A quantifier Q of type $\langle n_1, ..., n_k\rangle$ satisfies *domain independence* if and only if:

If $R_i \subseteq M^{n_i}$, $1 \leq i \leq k$, $M \subseteq M'$, then $Q_M(R_1, ..., R_k) \iff Q_{M'}(R_1, ..., R_k)$.

Informally, it means that the size of the universe does noes matter; if $Q_M$ is a $\langle 1, 1\rangle$ quantifier, $Q_M(A, B)$ depends only on the cardinality of A and the cardinality of B, and, importantly, not on the cardinality of M. So if, the cardinality of A and B remains the same, it does not matter what the cardinality of the other elements of M are. Compare (a) and (b) in Figure 2.1 for a visual representation of this difference. To exemplify, $Q_{three}(A, B)$ and $Q_{All}(A, B)$ do not depend on $|M - A \cup B|$, but only on the elements of A that are B, and so they are both extensional. Even ‘first three’ and ‘every third’ above have this property. In fact, no natural language determiner seems to be non-extensional. In order to find such quantifiers, we need to venture into the realms of logic, where e.g. $\forall$ violates domain independence. For this reason, domain independence seems to be a universal property of natural language quantifiers.

Another prime candidate for such universality is *conservativity.* Formally, conservativity is defined as follows:

**Definition 2.1.8.** A type $\langle 1, 1\rangle$ quantifier is conservative if and only if for all M and all A, B $\subseteq$ M:

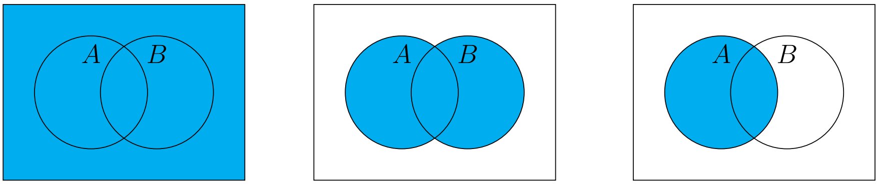
$Q_M(A, B) \iff Q_M(A, A \cap B)$

Informally, this is usually paraphrased as ‘Q As are As that are B’. It was recognized early on (Barwise & Cooper, 1981; Higginbotham & May, 1981; Keenan & Moss, 1985; Keenan & Stavi, 1986; van Benthem, 1984) that most determiners in natural language satisify this constraint. Consider the equivalence of sentences like:

18. (a) All dogs bark.

    (b) All dogs are dogs that bark.

19. (a) Three women dislike John.



(a) Domain dependent and non-conservative

(b) Domain independent and nonconservative

(c) Domain independent and conservative

Figure 2.1: Venn diagrams representing relevant objects for quantifiers with different logical properties.

(b) Three women are women that dislike John.

20. (a) Most people like Mary.

    (b) Most people are people that like Mary.

21. (a) Four out of five dentists recommend sugarless gum for their patients who chew gum.

    (b) Four out of five dentists are dentists who recommend sugarless gum for their patients who chew gum.

Conservative determiners consequently allow a further restriction of the universe, such that the truth value of a quantified statement $Q_M(A, B)$ with a conservative quantifier relies only on A. This is intuitive as well, as it reflects the primacy of the first argument of the quantifier; a claim like 'All dogs bark' is intuitively *about* dogs, not about things that bark (van Benthem, 1986a). Visually, this is illustrated in the Venn diagram in Figure 2.1 (c). Furthermore, van Benthem (1984) showed that only allowing conservative determiners, reduces the number of possible determiners considerably (from $2^{4^n}$ to $2^{3^n}$, where $n$ is the number of elements in the universe), and Keenan and Stavi (1986) showed that all conservative determiners were denotable by an English determiner, thus indicating that conservativity is an important property for natural language quantifiers.

However, the claim that it is universal has been put under pressure ever since its conception by apparent counterexamples:

22. Many Scandinavians have won the Nobel Prize in literature. (Westerståhl, 1985)

23. Mostly men walk. (Johnsen, 1987)

24. Only willows weep. (de Mey, 1991)

25. The company hired 75 % women. (Ahn & Sauerland, 2017)

26. The company hired all women. (Zuber & Keenan, 2019)

There is an evident discrepancy between these quantifiers and the conservative quantifiers above, in that rewriting them in $Q_M(A, A \cap B)$ form does not result in truth conditionally equivalent statements:

27. Many Scandinavians are Scandinavians who have won the Nobel Prize in literature.

28. Mostly men are men who walk.

29. Only willows are willows that weep.

This is because, as discussed in the context of value judgement quantifiers, 'many' can have a reverse proportional reading ($\frac{|A \cap B|}{|B|} > d$). Westerståhl (1985) pointed out that this reading is non-conservative; in particular, it depends on the Bs that are not A, which are not in the intersection of A and B, and so do the other two. 29 is definitely a tautology, and, unless strengthened by a scalar implicature *not all*, so is is 28. Such rephrasings are not available for so-called *bare proportionals* (Zuber & Keenan, 2019), as these do not appear in the subject position. They are, however, non-conservative, as their truth-value depends not on the proportion of the quantified noun phrase, but on the proportion of the other argument (in this case company hires).

Interestingly, these non-conservatives are the inverse of other conservative quantifiers, or have an equivalent sentence where the order of the arguments is reversed:

30. Many Nobel Prize winners in literature are Scandinavians.

31. Most walkers are men.

32. Every weeper is a willow.

33. 75 % of company hires were women.

34. All company hires were women.

This suggest that the alleged counterexamples are systematically related to conservative quantifiers by the inversion relation. Zuber (2004) examines a class of Polish quantifiers that translate into English 'only', and shows that only determiners that are the inversion of a conservative quantifier are grammatical determiners. This has lead people to explore the possibility that quantifiers may be conservative on either argument (see e.g. von Fintel & Matthewson, 2008, and references therein).

In the past, semanticists have referred to independent evidence to suggest that seemingly non-conservative determiners, like 'many', 'mostly' and 'only', are in fact not determiners, but adverbs or adjectives (e.g. Keenan & Stavi, 1986; van Benthem, 1986b; von Fintel, 1997). Recently, however, in light of the ubiquitousness of non-conservative expressions cross-linguistically (e.g. Ahn & Sauerland, 2017; Keenan & Paperno, 2017), an effort has been made to explain or salvage the conservativity universal (e.g. Romero, 2015; Romoli, 2015; von Fintel & Keenan, 2018; Zuber & Keenan, 2019) by weakening it, reanalyzing the alleged non-conservative quantifiers, or deriving the conservativity from other aspects of the linguistic structure. The latter solution is particularly pertinent, given that studies of the learnability of quantifiers have shown that conservative quantifiers have different learnability properties, compared to for example domain independence and monotonicity (Steinert-Threlkeld & Szymanik, 2019; van de Pol et al., 2019). However, I will remain agnostic as to the universality of conservativity, as it suffices for what follows, that the relevant quantifiers are.

Conservative and domain independent quantifiers that are closed under isomorphism are known in the literature as *CE quantifiers*. It was hypothesized in the foundational literature (Barwise & Cooper, 1981; Keenan & Stavi, 1986) that all natural language determiners are CE quantifiers, and excluding the examples discussed above, this hypothesis has been largely corroborated cross-linguistically (Keenan & Paperno, 2017). It might seem puzzling, given the diversity of the world's languages and the expressive power that these language exhibit, that determiners should denote only CE quantifiers. While I will not go into the technicalities here (but see e.g. Keenan, 2002), there are many other mathematical properties of importance to natural language generally and generalized quantifiers specifically – such as Boolean closure – that converge to yield precisely this set of quantifiers. Naturally, this leads to CE quantifiers having properties that are highly interesting for understanding language, and even cognition.

## 2.2 Semantic Automata

Among the interesting properties of CE quantifiers, is that since they are concerned with only one set of objects that either has or does not have a certain property, they can be represented by strings of binaries. From such strings one can pose computational problems that can be solved by abstract machines, or *automata*, thereby discovering what computational resources are minimally required to verify a quantified expression. Recall from 1.1, the intuitive idea that sets of strings can be used to represent Barwise and Cooper's (1981) "families of sets" that make a quantified expression true, argued to be the denotation of generalized quantifiers. The abstract machines described below are consequently charged with the task of determining whether a string belongs to the set of true strings – i.e. true models of a quantified expression – which is equivalent to a verification procedure.

The idea to construct verification algorithms originated with van Benthem's (1986c) seminal 'Semantic Automata', in which several interesting computational properties of quantifiers were proven. As mentioned in chapter 1, this specific way of characterizing quantifiers leads to precise complexity predictions that can be empirically tested. I will turn to these and subsequent classification results in 2.3, but I will first discuss the mathematical nature of these automata and the formal languages associated with them in 2.2.1 and how these relate to the verification of quantifiers in 2.2.2.

### 2.2.1 Automata Theory and their Corresponding Formal Languages

Before applying automata theory to quantifiers, I will first provide formal definitions of certain relevant automata. These definitions rely heavily on the canonical definitions in Hopcroft and Ullman (1979).

**Definition 2.2.1.** An alphabet, $\Sigma$, is a finite set of symbols. A string, $w$, is a finite sequence of symbols. The length of a string, $|w|$, is the number of symbols in the string. If $|w| = 0$, then w = $\epsilon$, *the empty string. Concatenation* of $w_0$ and $w_1$ means joining $w_0$ and $w_1$ together to form a new string $w_0w_1$. The concatenation of $w$ and $\epsilon$ is the same as $w$: $\epsilon w = w\epsilon = w$.

**Definition 2.2.2.** A *finite state automaton* (FSA) is 5-tuple $(Q, \Sigma, \delta, q_0, F)$, where Q is a finite set of states, $\Sigma$ is a finite input alphabet, $q_0$ is the initial state, $F \subseteq Q$ is the set of final states, and $\delta$ is a function from the Cartesian product $Q \times \Sigma$ to $Q$, called the *transition function.*

Less formally, an FSA is an abstract machine that can be in various states. It begins in the starting state, and then reads a string of symbols one at a time. For each symbol, the transition function determines whether the machine will stay in the same state or move to another state, depending on the input and the state it is currently in. The set of final states, $F$, are accepting states. If the machine is in one of these states when it finishes reading the input, we say that the FSA recognizes, or accepts, this string. If it is in any other state, i.e. not an accepting state, the string is not recognized by the FSA. See Figure 2.2 for a visual representation of such a machine.

The set of strings recognized by an FSA, and other such abstract machines, is called a *formal language.*

**Definition 2.2.3.** A formal language, L, is a set of strings from one alphabet. The empty set, $\emptyset$, and the set of the empty string, $\{\epsilon\}$ are languages. The language of all strings over an alphabet $\Sigma$ is denoted $\Sigma^*$, such that if $\Sigma = \{a\}$, then $\Sigma^* =$

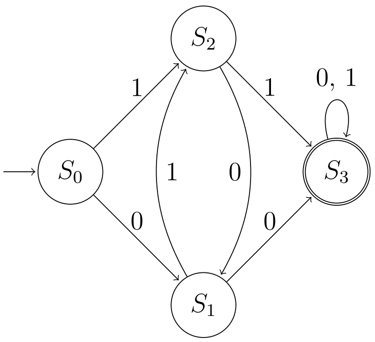


Figure 2.2: Example of a finite state automaton. States $S_0$ through $S_3$ are represented by the labeled circles. The initial state, indicated by an arrow from nowhere, is $S_0$. The transition function is represented by the arrows between states; the label on the arrow indicates the string input, and the arrowhead indicates the direction of the transition. For example, if the machine is in state $S_0$ and reads a 1, it transitions into state $S_2$. If it reads a 0, it goes to state $S_1$. Staying in the same state is indicated by a looping arrow, as seen in $S_3$, where the automaton stays in the same state, regardless of the input. Accepting states are indicated by a double-lined circle. In this case $S_3$ is the only accepting state. This automaton in particular recognizes all and only those strings that contain either two consecutive 1s or two consecutive 0s.

$\{\epsilon, a, aa, aaa, ...\}$. For all sets of strings $L, L_1, L_2, ...$ from $\Sigma^*$, let $L^0 = \{\epsilon\}$ and $L^i = LL^{i-1}$ for $i \geq 1$.

The *Kleene closure* of L, $L^*$, is the set:
$L^* = \bigcup_{i=0}^{\infty} L^i$
and the *positive closure* of L, $L^+$, is:
$L^+ = \bigcup_{i=1}^{\infty} L^i = L^* - \epsilon$

A foundational proof from formal language theory (Kleene, 1951), demonstrated that the set of strings recognized by finite state automata are precisely the sets that belong to *regular languages.*

**Definition 2.2.4.** For an alphabet $\Sigma$, a *regular language* over $\Sigma$ is recursively defined:

1. The empty language $\emptyset$ is a regular language.

2. For each $a \in \Sigma$, the singleton language $\{a\}$ is a regular language.

3. If $A$ and $B$ are regular languages, the union $A \cup B$ and concatination $AB$ of A and B are also regular languages.

4. For a regular language $A$, $A^*$, and consequently $\{\epsilon\}$, is a regular language.

5. No other languages over $\Sigma$ are regular.

*Regular expressions* are a convenient way to describe these languages, and are defined recursively in a similar fashion:

**Definition 2.2.5.** For an alphabet $\Sigma$:

1. $\emptyset$ is a regular expression and denotes the empty set.

2. $\epsilon$ is a regular expression and denotes the set $\{\epsilon\}$.

3. For every symbol $a$ in $\Sigma$, $a$ is a regular expression denoting the set $\{a\}$.

4. If $r$ and $s$ are regular expressions denoting the respective languages R and S, $r + s$, $rs$, and $r^*$ are regular expressions denoting $R \cup S$, RS, and $R^*$, respectively.

By convention, $*$ has higher precedence than concatenation, which in turn has precedence over $+$.

The set of regular languages is thus highly restricted compared to all possible formal languages, and constitutes the innermost, or bottom, level of the so-called *Chomsky hierarchy* (Chomsky, 1956). Many facets of language, such as phonological patterns and most morphological systems, are in fact regular (e.g. Gazdar & Pullum, 1985; Kaplan & Kay, 1994; Langendoen, 1981). It was argued early on, that syntax was not regular (Chomsky, 1956), but has to be *context free*. *Context free languages* are more complex and more expressive than regular languages, and are defined by reference to *context free grammars*, and I will define both in what follows.

**Definition 2.2.6.** A *context free grammar* $G = (V, P, T, S)$, where V and T are disjoint finite sets of variables and terminals. P is a finite set of productions of the form $A \to \alpha$, where A is a variable and $\alpha$ is a string of symbols from $(V \cup T)^*$. S is a special variable called the *start symbol*.

If $\alpha$ and $\beta$ are strings of variables, $\alpha$ *derives* $\beta$ if $\beta$ follows from $\alpha$ by zero or more productions of P.

**Definition 2.2.7.** The *language* generated by G is denoted L(G). A string $w$ is in L(G) if $w$ can be derived from $S$ and $w \in T^*$.

A language L, is *context free* if it is L(G) for some CFG $G$. Two grammars $G_1$ and $G_2$ are equivalent if $L(G_1) = L(G_2)$.

The canonical example of a context free language, $a^n b^n$, where $n$ number of $a$s are followed by exactly $n$ $b$s for any $n$, can be generated by a CFG $G$, where $V = \{S\}, T = \{a, b\}, P = \{S \to aSb, S \to ab\}$. Applying $S \to aSb$ $n-1$ times, followed by one application of $S \rightarrow ab$, yields $a^nb^n$. Since $a^nb^n$ can be derived from $S$ and consists only of $a$s and $b$s – i.e. only terminals – $a^nb^n$ is a context free language.

Just like regular expressions have an automaton equivalent, context free grammars have an automaton counterpart in *pushdown automata.* Essentially, a pushdown automaton (PDA) is an FSA augmented by a memory stack, where information about the previous input can be stored. More formally:

**Definition 2.2.8.** A *pushdown automaton* (PDA) is defined as a 7-tuplet $(Q, \Sigma, \Gamma, \delta, q_0, Z_0, F)$, where Q is a finite set of states, $\Sigma$ is the input alphabet, $\Gamma$ is the stack alphabet, $q_0 \in Q$ is the initial state $Z_0 \in \Gamma$ is the start symbol, $F \subseteq Q$ is the set of final states, and $\delta$ is a mapping from $Q \times (\Sigma \cup \{\epsilon\}) \times \Gamma$ to finite subsets of $Q \times \Gamma^*$.

This is necessary to account for languages such as $a^nb^n$ because in order to know whether there are as many $a$s as $b$s for any arbitrary $n$, it is not sufficient to transition into a potential accepting state on the basis of the present input. Let us say that the present input is $b$ and the automaton is in the state where it would be if the previous input was three consecutive $b$s. But this state contains no information about the number of previous $a$s, since it came from a state that contained a $b$ and is now seeing a $b$, and so the automaton cannot determine if there are as many $b$s as $a$s unless this information is stored elsewhere in the machine, i.e. in a memory stack.

Because of the addition of a memory stack, the transition function is augmented, so that it is a function from a state, input, stack triplet to a state, stack pair. Intuitively, this means that the automaton considers its input, and depending on the state it is in and the topmost symbol on the stack, it stays in the same state or transitions to another state. Additionally, the transition function can simultaneously manipulate the stack by adding an element to the stack, popping off the topmost symbol, or not change the stack at all. See Figure 2.3 for an example.

There are two more complex layers in the Chomsky hierarchy that contains the *context sensitive* and *recursively enumerable* formal languages with their corresponding automata, the *linear bounded automaton* and the *Turing machine.* As these higher tiers of the Chomsky hiearchy are not relevant for the discussion to follow, I will not describe them in any detail, but it should be mentioned that some people have presented linguistic evidence to suggest that syntactic dependencies are not context free, but mildly context sensitive (Shieber, 1985). Since inner, or lower, layers are recognized by outer, or higher, levels in the Chomsky hierarchy, this has been used to argue that the complexity of the linguistic computational procedure is determined by the most complex expressions that it is possible to denote. So even though some syntactic expressions are in fact regular, if the most complex syntactic expressions are mildly context sensitive, even the regular dependencies should be de-

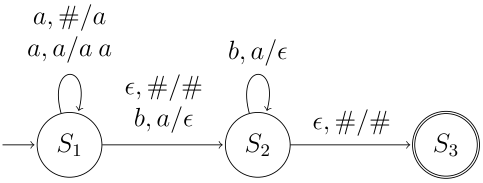


Figure 2.3: Example of a PDA recognizing the language $a^n b^n$, i.e. an equal number of consecutive $a$s and $b$s for any $n \geq 0$. Intuitively, the machine reads all the $a$s and stores them on the stack, before popping off one $a$ for each $b$ it sees. If there are as many $b$s as $a$s, the stack memory will be empty and the machine can proceed to the accepting state. More specifically, the automaton starts in the initial state $S_1$ with the start symbol # on the stack. If it reads an $a$ as the first character it sees – i.e. the stack is empty, as represented by the start symbol – it pushes an $a$ to the top of the stack. If it reads an $a$ and the topmost symbol on the stack is also $a$, it puts another $a$ atop the stack. If it reads a $b$ and there is an $a$ at the top of the stack, it pops off the topmost $a$ and proceeds to $S_2$. It stays in $S_2$, popping off $a$s from the stack, as long as it reads $b$s. When there are no more characters, i.e. it reads $\epsilon$, it proceeds to the accepting state $S_3$ as long as the string is empty, i.e. the symbol at the top of the stack is the start symbol #. This automaton also recognizes the empty string $\epsilon$ ($a^0b^0$), since when the string is the empty string, the stack is also empty, and the automaton can immediately continue to $S_2$ and the accepting state $S_3$. Notice that there are many options that are not accounted for in this model: If, e.g., the first symbol on the string is $b$, or # is not atop the stack when the machine has read the last $b$, there are no legal actions described by the transition function. In this case, the automaton stops and does not reach the accepting state.

scribed as mildly context sensitive. Also worth noting is a growing body of literature that has shown that phonological patterns are not only regular, but subregular, i.e. constitutes a proper subset of the regular languages (Heinz, 2018; Heinz & Rawal, 2011). Related work also suggests that morphological patterns (Chandlee, 2017), syntactic dependencies (Graf, 2012, 2017), and, to a certain extent, lexicalized determiner semantics (Graf, 2019) can be described in subregular terms, given certain assumptions. If this is the case, it could potentially lead to a unified account of the computational properties of the language faculty. However, I will set this aside for now, and return to how these results can inform the work herein in chapter 6.

At this point it might be poignant to address what measure of complexity is relevant. One approach, taken by, e.g., Ristad (1993), is to define complexity in terms of computing time and space, i.e. the duration of an algorithm and/or the amount of memory required to perform a computation, respectively. This approach has also been applied to quantifier semantics (Szymanik, 2010), but in the present context, we are more concerned with *expressivity*, also called *expressive power*: what expressive power is needed to define the set of strings, i.e. models, that make a quantifier true?[4] We can define expressivity over strings, as the ability to define a larger set of sets of strings. For example, it has been shown that first-order logic is a subset of the regular languages (more precisely the star free languages) (McNaughton & Papert, 1971), i.e. all the sets of strings that are definable in first-order logic are regular, but not all regular languages are definable in first-order logic. In essence, this is what the Chomsky hierarchy shows: while FSAs and regular expressions have exactly the same expressive power, context free grammars and PDAs can describe a strictly larger set of string languages (Chomsky, 1956).

Chomsky (1965) argued that it is only when some grammar fails to generate the language one is attempting to characterize that *expressivity* is useful to linguistic theory, because the reason it fails can tell us about essential properties of the language that distinguishes it from other languages of different complexity (see also De Santo & Rawski, 2022). However, the cognitive correlates of expressivity are more opaque, at least on a general level, and require explicit linking hypotheses. I will return to the specific linking hypotheses for this thesis in 2.4 and 3.3, and discuss the general problem of linking differences in expressive power to cognitive effects in chapter 6.

### 2.2.2 Algorithms of Quantifier Verification

As mentioned at the beginning of this section, it is possible to code CE-quantifiers as strings of binary input. Informally described, the algorithm associated with a quantifier $Q_M(A,B)$ is fed a list with the elements of A. All elements are either 0 or 1, where 0 are elements that are A but not B ($a \in A - B$) and 1 are elements that are in both A and B ($a \in A \cap B$):

**Definition 2.2.9.** A quantifier corresponds to a class Q, where Q is a language $L_Q$, describing all models of Q.

To construct CE quantifier languages, one thus enumerates the set A for a quantifier $Q_M(A,B)$, and writes a 0 for every element in $A - B$ and a 1 for every element in $A \cap B$.

It is also possible to have a four symbol alphabet – e.g. 0, 1, 2, 3 – where each symbol denotes membership of A or B. Then 0 could be not A and not B, 1 be A but not B, 2 be B but not A, and 3 be both A and B. This could also be generalized to quantifiers that relates more than two sets (e.g. M. Mostowski, 1998):

**Definition 2.2.10.** The class Q can be represented by the language $L_Q$, that does not contatain the empty string, over the alphabet $A = \{\alpha_0, ...\alpha_{2^n-1}\}$, such that $\alpha \in L_Q$ if and only if there is $(U, A_1, ..., A_n) \in Q$ and a linear ordering of $U =$

[4] See e.g. Gazdar and Pullum (1985) for an application of this notion of complexity to syntactic patterns.

$(b_1, ..., b_k)$, where $k$ is the length of $\alpha$ and the $i$-th character of $\alpha$ is $a_j$ exactly when $b_i \in S_1 \cap ... \cap S_n$, where:

$$S_l = \begin{cases} A_l \text{ if interger part of} \frac{j}{2^{l-1}} \text{ is odd} \\ U - A_l \text{ otherwise} \end{cases}$$

However, since we are only concerned with CE quantifiers, we will proceed using binary notation.

These definitions parallel Barwise and Cooper's (1981) notion that quantifiers denote families of sets. In the space of all possible sets of a model, a quantifier can be seen as dividing up the sets on the basis of which sets are true and false of the quantifier. Every quantifier can therefore be said to denote all the sets that makes the quantifier true. Similarly, we have defined a quantifier as a class of models that make the quantifier true, and a quantifier can denote all and only those strings that represent these models. Since such sets of strings are, by definition, formal languages, every CE quantifier corresponds to a formal language over a binary alphabet.

**Definition 2.2.11.** Let Q be a type $\langle 1, 1 \rangle$ quantifier. Then, *the language of* Q is defined as:

$$L_Q = \{s \in \{0,1\}^* | (\#_0(s), \#_1(s)) \in Q\}$$

where $\#_a(\alpha)$ is the number of occurences of a symbol $a$ in a word $\alpha$.

Because formal languages correspond to automata, this means that every quantifier is not only associated with a formal language, but is also associated with a corresponding automaton that recognizes that language. Considering that the language of a quantifier Q consists of all the strings representing true models of Q, this is effectively a verification algorithm. As described in 2.2.1 above, languages correspond to automata of varying degrees of complexity, and I will now go on to show that natural language quantifiers can be classified on the basis of the computational resources required to verify them (M. Mostowski, 1998; Szymanik, 2016; van Benthem, 1986c).

## 2.3 Quantifiers Classified by Algorithmic Complexity

Recall that Aristotelian quantifiers denote the following relations between sets A and B:

$$Q_{all} A, B \iff |A \cap B| = |A|$$

$$Q_{some} A, B \iff |A \cap B| \neq 0$$

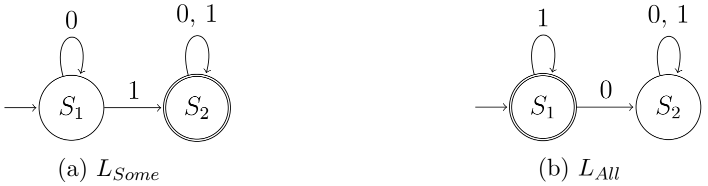


Figure 2.4: Aristotelian quantifier automata

$$Q_{no}A, B \iff |A \cap B| = 0$$

$$Q_{not\,all}A, B \iff |A \cap B| \neq |A|$$

As is apparent, these fall into two kinds: 'all' and 'no' require all As to be of one type – either $|A \cap B|$ or $|A - B|$ – whereas 'some' and 'not all' requires at least one instantiation of one of the types. Translating these into string languages, 'all' corresponds to the language $L_{All} = 1^+$, meaning it consists of all languages containing only 1s. 'Some' corresponds to the language $L_{Some} = (0+1)^*1(0+1)^*$: the language of strings containing at least one 1. Comparable languages can be constructed for 'no' and 'not all' by replacing the 1s with 0s.

Constructing automata for Aristotelian quantifiers is therefore fairly straightforward. The automata reads the string representing the model, and in the case of 'Some' it starts in the initial state $S_1$ and stays there until it sees a 1, upon which it changes into the accepting state $S_2$ (see Figure 2.4 (a)). If no 1 is found, the string is not accepted, and the sentence is false. For 'All' (see Figure 2.4 (b)), $S_1$ is both the initial state and the accepting state; it stays in this state as long as it reads 1, and if it reads a 0, it moves to $S_2$, which is not an accepting state. As a consequence, it only accepts models where all the As are Bs. As an illustration of what these machines are mathematically, consider how $FSA_{Some}$ can be expressed more formally:[5]

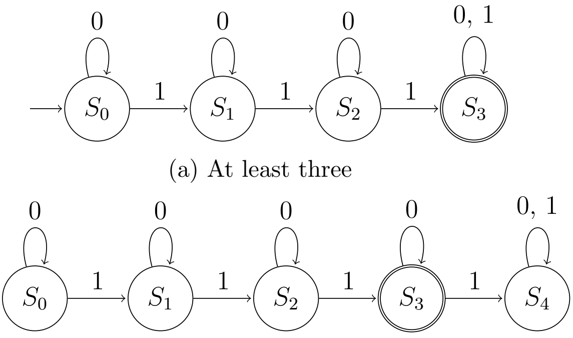
$$FSA_{Some} = (\{S_1, S_2\}, \{0, 1\}, \{((S_1, 0), S_1)), ((S_1, 1), S_2), ((S_2, 0), S_2), \\ ((S_2, 1), S_2)\}, \{S_1\}, \{S_2\})$$

Turning to *numerical* quantifiers, these set requirements on the precise cardinality of $|A \cap B|$. For any $Q_n$, $|A \cap B| = n$. As mentioned, numerical quantifiers give rise to an *exact* reading and an *at least* reading (e.g. Horn, 1972; Levinson, 1983):

$$Q_{exactly\,three}A, B \iff |A \cap B| = 3$$

$$Q_{at\,least\,three}A, B \iff |A \cap B| \geq 3$$

[5] Since this is notationally tedious, other quantifiers will not be expressed in this manner.



(a) At least three

(b) Exactly three

Figure 2.5: Numerical quantifier automata

These correspond to slightly different string languages: $L_{exactly\, three} = 0^*10^*10^*10^*$ and $L_{at\, least\, three} = (0+1)^*1(0+1)^*1(0+1)^*1(0+1)^*$. Both languages contain three 1s, but on the exact reading these three ones can only be surrounded by zero or more 0s, whereas on the at least reading they can be preceded and followed by zero or more combinations of 0s and 1s.

For both readings the automaton starts in the initial state $S_0$, and for all states up until state $S_n$ – in this case $S_3$ – it stays in the same state if it reads a 0, and changes to the next state if it reads a 1. $S_n$ is the accepting state, and on the *at least* reading, the automaton stays in this state regardless of whether it reads a 0 or a 1. On the *exactly* reading, however, it transitions into another state, $S_{n+1}$, that is not an accepting state, if it reads a 1. This way, the *exactly* automaton recognizes only models where exactly three As are B, whereas the *at least* automaton accepts all models where three or more As are B (see Figure 2.5).

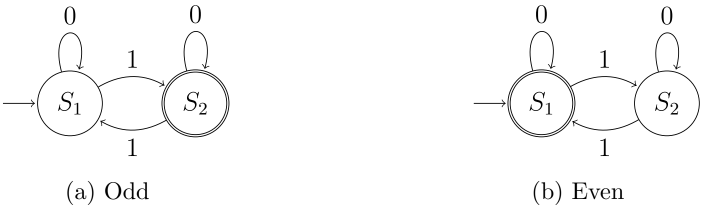
Turning next to the *parity* quantifiers – i.e. 'an even/odd number of' – recall that these require the cardinality of $|A \cap B|$ to be either even or odd – or equivalently, two divides or does not divide the cardinality of $|A \cap B|$. Consequently, the string languages associated with parity quantifiers are of the following form:

$$L_{odd} = 0^*10^*(0^*10^*10^*)^*$$

$$L_{even} = 0^*(0^*10^*10^*)^*$$

For odd numbers, every string begins with an arbitrary number of 0s followed by a 1 and an another arbitrary number of 0s, before zero or more iterations of two 1s with zero or more 0s at either side of them. The language for even numbers is similar, but the first string contains only zero or more iterations of 0.

Importantly, these quantifiers are not definable in first-order logic, and consequently they are not accepted by acyclic automata (van Benthem, 1986c). However, one can easily construct a cyclic automaton that recognizes $L_{odd}$ and $L_{even}$



(a) Odd

(b) Even

Figure 2.6: Parity quantifier automata

(M. Mostowski, 1998). Both automata start in $S_1$, reading the string of 0s and 1s, and change states every time they see a 1 (see Figure 2.6). Consequently, $S_1$ is the even state and $S_2$ is the odd state. As a consequence, the accepting state of $FSA_{odd}$ is therefore $S_2$, whereas $S_1$ is the accepting state of $FSA_{even}$.

All quantifiers considered thus far are therefore recognized by finite state automata. However, it is provable that *proportional* quantifiers are not recognizable by such automata, but require a pushdown automaton (PDA) to be computed (Kanazawa, 2013; van Benthem, 1986c). Recall that proportional quantifiers have the following semantics:

$$Q_{most}A, B \iff |A \cap B| \geq |A - B|$$

$$Q_{less\,than\,half}A, B \iff |A \cap B| \leq \frac{|A|}{2}$$

$$Q_{a\,third\,of}A, B \iff |A \cap B| = \frac{|A|}{3}$$

The string language associated with them, thus take the following form:

$L_{most} = 0^n 1^{n+k}$, where $k \geq 1$

$L_{less\,than\,half} = 0^{n+k} 1^n$, where $k \geq 1$

$L_{a\,third\,of} = 0^n 1^{\frac{n}{3}}$

Such languages are context-free. The corresponding PDAs – here illustrated with ‘most’ – can be constructed in the following fashion (see Figure 2.7): The automaton starts in the initial state $S_1$ with the start symbol # at the top of the stack. If the string is empty, it does nothing and stays in $S_1$, which is not an accepting state. If it reads a 1 or a 0, it puts this symbol at the top of the stack. Every time it encounters the opposite symbol to the one at the top of the stack – e.g. if the top of the stack is 1 and the input is 0 – it pops that symbol off the top of the stack, whereas if the input and the stack symbol is the same – e.g. 1 on the stack and 1 in the input – it puts the new symbol atop the other. When it reaches the end of the string, there are two possible scenarios. If 0 is the topmost symbol on the stack, it does nothing and stays in $S_1$, but if it is 1, it clears the stack and moves to the accepting state

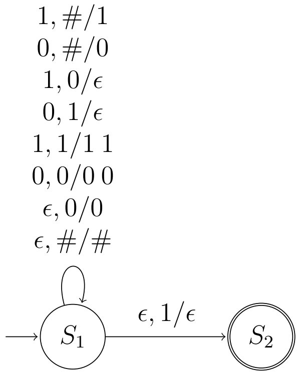


Figure 2.7: Proportional quantifier automaton recognizing 'most'

$S_2$. So the PDA only accept strings that represent models where the As that are B outnumber the As that are not B, which is the precisely the semantics for 'most' that we gave above.

## 2.4 Summary

This leads us to postulate the four complexity classes from the introduction:

| | | | |
|---|---|---|---|
| (1) | **Aristotelian**: | 'Some', 'all' | acyclic FSA |
| (2) | **Numerical**: | 'Three' | acyclic FSA |
| (3) | **Parity**: | 'An even/odd number of' | FSA |
| (4) | **Proportional**: | 'Most' | PDA |

Aristotelian and numerical quantifiers are both recognized by acyclic finite state automata. The difference between them is that the number of states in the Aristotelian quantifier FSAs are fixed, whereas the number of states, and consequently the complexity of the algorithm, increases proportionally to the $n$ denoted by the specific numerical quantifier for numerical FSAs. Parity quantifier automata are of the same kind as the Aristotelian quantifier automata, except that the FSA needs to be cyclic. Proportional quantifiers on the other hand, cannot be recognized by FSAs, but need a PDA with a memory component. The most significant difference is therefore between the proportional quantifiers and the other non-proportional quantifiers.

We have seen that natural language determiners correspond to type $\langle 1, 1 \rangle$ CE quantifiers that denote relations between sets of individuals. The denotation of such quantifiers can be said to be the family of sets that make the quantitative relations expressed by the quantifier true. These families can be represented by formal languages over a binary alphabet that are recognized by automata. Quantifiers fall into four classes depending on the computational expressivity of the associated automata, and the most significant difference is between non-proportional quantifiers, that can all be computed by finite state automata, and proportional quantifiers, that can only be computed by pushdown automata with a memory component.

Since the difference between FSAs and PDAs is the addition of a stack memory, the linking hypothesis between computational expressivity and cognitive resources is more straightforward than, e.g., the difference between acyclic and cyclic automata. Humans rely on memory of different kinds, such as working memory (Baddeley, 2012) or recollection memory (Yonelinas, 2002), which is modulated by task demands. Consequently, we expect the verification of proportional quantifiers to elicit neural signatures of increased memory load during sentence processing that are not found in the processing of non-proportional quantifiers.

# Chapter 3
# Methodological Considerations

We have seen that verification algorithms for natural language quantifiers delineate four distinct classes of quantifiers as a function of the computational resources involved in the computation. In particular, it is hypothesized that proportional quantifiers require additional memory resources compared to other kinds of quantifiers, and that this is reflected in the verification of quantified sentences in human subjects. The aim of this project is to assess whether such differences are detectable at the neural level, and it is consequently essential to describe how to measure the workings of the human brain and to discuss the potential caveats as well as the precautionary steps that can be taken to overcome these. The nature of these precautions is not uniform. On the one hand, theoretical – perhaps, even philosophical – considerations about what role neural data can play in explanations is important as a driving force for the project presented in this thesis. On the other, the impedance of this force by very tangible realities about the available methods for gathering the neural data, by necessity, plays an equally important role.

Brain activity is generated by cascades of chemical reactions causing changes in the electric polarization of the cell membrane of individual neurons. This, in turn, leads to the neurons transmitting, or inhibiting, signals to other neurons via its axonal connections. The neuronal activity – both of a single neuron and groups of neurons – are typically conceived of as the conduits of an information processing system, and such activity is therefore the object of study for cognitive neuroscience. However, while there are invasive studies that implant electrodes in animal brains to study the activity of individual neurons, in practice, this is not possible for experiments on humans. One is therefore relegated to studying the remnants of such activity detectable outside of the scalp, and this causes some methodological problems that are the first topic of this chapter. After justifying the choice of electroencephalography (EEG) in 3.1, I will deal more specifically with the challenges related to this technique, as well as with the importance of experiments with a strong theoretical foundation in 3.2. The relevance and importance of algorithmic analysis is the topic of 3.3, and section 3.4 is devoted to synthesizing the preceding discussion into concrete considerations of the experimental design. Finally, I will provide a note on how Norwegian quantifiers relate to quantification in languages that might be more familiar to the reader in 3.5. A summary of the chapter is found in 3.6.

## 3.1 Picking your Poison: Advantages and disadvantages of different neuroimaging techniques

As mentioned above, it is seldom possible to study the activity of individual neurons in humans. However, technological advancements – largely in computer processing power – in recent years have seen the advent of ever more refined functional neuroimaging techniques that try to circumvent this limitation. Largely speaking, there are two main approaches: (1) one can either study brain activity indirectly, by observing changes in blood-flow and oxygen consumption in different areas of the brain, or (2) one can directly measure the distorted electrical activity reflected in relative changes in the electrical potential at the scalp. Both of these approaches are used extensively in neurolinguistics (see, e.g., the respective chapters of de Groot & Hagoort, 2018), and both have advantages and disadvantages, predominantly manifested in a tradeoff between spatial and temporal resolution. I will address these approaches in 3.1.1 and 3.1.2, respectively, before justifying my decision to choose the second approach in 3.1.3.

### 3.1.1 Arsenic: The BOLD response

*Blood-oxygen level dependent* imaging is an indirect measure of brain activity that relies on the fact that increased brain activity is associated with an increase in blood-flow and oxygen consumption. Since neuronal firing requires energy, a higher firing rate causes oxygenated blood to flow to the brain areas with increased activity. Therefore, measuring differences in blood-flow can inform us about the loci of processing in the brain by contrasting the BOLD response between experimental conditions.

The imaging techniques associated with this measure is *functional Magnetic Resonance Imaging* (fMRI) and *functional Near-Infrared Spectroscopy* (fNIRS). The most important part of MRI equipment is a large magnet, and it records disturbances in the magnetic field created by this magnet. fMRI therefore relies on the different magnetic properties of oxygenated and deoxygenated hemoglobin – oxygenated hemoglobin being diamagnetic and deoxygenated being paramagnetic – to measure increases in blood-flow and oxygenconsumption of populations of nerve cells in different areas of the brain. fNIRS, by contrast, emits near-infrared light into the head of a subject, and measures the fracturing and absorption of the light as it leaves the scalp. In this case, it is the different properties of near-infrared light absorption of oxygenated and deoxygenated hemoglobin that is the dependent measure. However, since it relies on light absorption, it is not able to penetrate as deeply into the brain as, e.g., fMRI, but can only detect changes in outer layers of the cortex.

While the spatial resolution of these techniques are impeccable, their temporal resolution is impeded by the hemodynamic response: canonically, it is assumed that there is a 2 second delay period between a stimulus and the onset of the hemodynamic response, that it peaks after about 6 seconds, and that it sustains until the stimulus disappears (Buxton, 2009). The response also saturates over time, meaning that the BOLD response to a second stimulus of the same type, if not time-shifted appropriately, is smaller than that to the initial stimulus (Buxton, Uludağ, Dubowitz, & Liu, 2004; Lindquist, Loh, Atlas, & Wager, 2009). It has been suggested (Polimeni & Lewis, 2021), that more rapid fMRI sampling rates with more fine-grained resolution of the hemodynamic response, e.g. vessel dilation, promise the measurement of much faster brain dynamics, but the practical application of such techniques are largely still in the future.

Hemodynamic approaches are consequently best tailored to answer where-questions. Whenever when-questions are in order, electrophysiological measures are more apt.

### 3.1.2 Cyanide: Electrophysiological measures

As mentioned in the introduction, neuronal activity is electric. When large populations of neurons fire at the same time, the sum of the dipoles generated by the current flowing from the cell body to the apical dendrites, generates electrical currents large enough to create fluctuations in the brain's electric field. If, additionally, the orientation of the neuronal dipoles are aligned and radial to the skull, the activity can be detected by amplifying the signal from electrodes placed on the scalp (e.g. Luck, 2014). This technique is known as electroencephalography (EEG).

This has the advantage that you can measure neuronal activity directly and in real time, and therefore does not suffer from the poor temporal resolution of methods reliant on the BOLD signal. However, extracranial EEG suffers from two electrophysical properties of the head: (1) the brain is conductive, viz. the electricity spreads out as it passes through the brain, and (2) bone has high electrical resistance, meaning that the currents that do reach the scull are diverted, causing further spatial blurring. For this reason, the spatial resolution of EEG is generally quite poor. Source localization of the dipole generators can only be done through hypothetico-deductive methods where a generator location is assumed, since there are infinitely many possible dipole configurations that can explain an observed voltage distribution.

There are certain ways to circumvent some of these limitations. Intracranial

EEG removes the further spatial blurring of the scull, but is very invasive and is thus usually reserved for medical purposes. Magnetoencephalography (MEG) relies on measuring the magnetic field generated by any current, and since magnetic fields are not disturbed by electrical resistivity, the location of the currents can be determined with a higher degree of certainty. However, MEG does not overcome the problem of spread through the conductive medium, nor does it detect all the currents. Because the magnetic field revolves around a current, only dipoles that are tangential to the surface of the scalp can be detected, whereas the perpendicular dipoles, which are the ones measured by EEG, never leave the head, and consequently cannot be measured. MEG and EEG therefore measure complimentary dipoles. Additionally, the technology used in MEG is reliant on superconductivity, which means cooling the probes using liquid helium. This makes MEG equipment very expensive to use, even setting aside the cost of the equipment itself.

### 3.1.3 Hemlock: The means to answer the research questions

When picking a poison, the nature of the effects one is expecting is therefore of the essence. The precise temporal localization of neuronal activity is best done with electrophysical methods, whereas the involvement of specific brain regions is more easily achieved using hemodynamic measures. Additionally, practical trivialities, such as time-constraints, cost, and the availability of equipment, play a role.

The effects we are looking for in this project, are predominantly those of memory. Such effects are typically prolonged in time, and could therefore in principle be detected even with the low temporal resolution of BOLD response methods. In fact, there is preliminary fMRI evidence, to be discussed below, of the involvement of memory systems in quantifier verification (McMillan et al., 2005; Olm et al., 2014). An fNIRS study to try to replicate these results was planned as a part of this project, but time-constraint imposed by unforeseen circumstances, predominantly of a pandemic nature, made this untenable. Regardless, ideal answers to the research questions should expose both the brain regions involved in as well as the temporal order of the various stages of verification. While the brain regions involved will have to be inferred from the literature – see paper 3 – the impact of verificational complexity on the different stages of sentence processing is also unknown.

We understand the meanings of words and sentences very rapidly, and the different stages of sentence processing are therefore measured in milliseconds, way below the several seconds required to elicit a BOLD response. Since one of the research questions is how the computational complexity of verification affects sentence processing, it was necessary to implement electrophysiological measures. MEG equipment was not available, and consequently, EEG was used to this end.

In analyzing EEG results, there are a number of options as to which aspects of these multifaceted data to utilize. The most widely used analyses are *time-frequency representations* (TFRs) and *event-related potentials* (ERPs). TFRs emphasize the oscillatory activity of neurons, i.e. rhytmic patterns of neuronal firing. The synchronization and desynchronization of such activity in different regions of the cortex, is thought to reflect coupling and decoupling of the various functional networks involved in cognitive processing (Bastiaansen, Mazaheri, & Jensen, 2012). ERPs, by contrast, are amplitude changes evoked by some *event* that it is time-locked to. Since various cognitive processes are active at any given time – as evidenced by the constant fluctuations in the EEG signal – ERPs seek to eliminate the impact of the *background EEG*, i.e. random variation in the EEG signal, by averaging over a large number of trials, so that the cogntive process of interest is isolated. In this averaged signal, one finds patterns of positive and negative deflections in the waveforms that vary systematically with, and thus plausibly index, functional processes (Rommers & Federmeier, 2018). Waveforms with established links to cognitive processes are traditionally labelled *ERP components*.

There has been increasing interest in TFRs in the last decades, and of particular interest here, is the systematic correlations between frequency band power and memory systems that have been established (for a review, see Lisman & Jensen, 2013). TFR analysis would therefore be informative as to whether memory systems are involved in the verification of quantifiers. However, while the knowledge of the oscillatory dynamics of sentence processing has recently become quite considerable (Meyer, 2018), the corresponding ERP literature is much more established (Swaab, Ledoux, Camblin, & Boudewyn, 2012). Since this is the first systematic study of the impact of computational complexity on quantifier verification using EEG, and the nature of the effects was consequently unknown, a more traditional approach also has merit. For these reasons, ERP analysis was deemed appropriate for the studies presented here, but future research should look into possible effects in the time-frequency domain as well.

## 3.2 The Aims of Cognitive Science and the Kinds of Experiments they Necessitate

One issue with neuroimaging in general, that is particularly prevalent with event related potentials (ERPs), is what role such data play in explanations. If "cognitive neuroscience is the scientific study of how neural activity explains cognition and the behavior it gives rise to" (Boone & Piccinini, 2016, p. 1515), then it is not immediately obvious what the explantory role of ERPs are. In particular, the various ERP components are defined on the basis of which cognitive processes manipulate them. This is particularly evident for components such as the *Mismatch Negativity*

(MMN) or the *Error-Related Negativty* (ERN) – reflecting change detection in auditory stimulation (Näätänen & Kreegipuu, 2012) and the awareness of an incorrect response (Gehring, Liu, Orr, & Carp, 2012), respectively – but it is also the case for more generically labeled components such as the P300, which is associated with increased attentional focus (Polich, 2012).

Effectively, this is a reversal of the *explanandum* – the phenomenon to be explained – and the *explanans* – its explanation – regardless of the specific notion of explanation you deploy. While a full review of what consitutes a scientific explanation is beyond the scope of this chapter (but see e.g. Salmon, 1989; Weber, Van Bouwel, & De Vreese, 2013), a brief summary of some key terminology and insights from this literature is necessary. Such a summary is provided in 3.2.1, and a derivation of how these insights should inform ERP research is found in 3.2.2.

### 3.2.1 Notions of Explanation

Firstly, explanations are answers to why-questions (Garfinkel, 1981; Lipton, 1991, 2004; van Fraassen, 1977, 1980). This means that explanations are inherently interest relative, in that they depend on the question asked. For example, the question “why is this carrot rotten?” can be answered by “because it’s been in the fridge for ages” or “because at this stage of biodegeneration, the aerobic digestion of microorganisms has caused the polymer of the carrot to decompose into oligomers and monomers”, depending on whether it is an informal inquiry about a particular carrot, or a generic question about the chemistry of rot. Moreover, explanations are *contrastive*, since why-questions can, usually, be paraphrased as “why P rather than Q” (Lipton, 2004). As an example, consider the question “why is this carrot rotten?” again. If the question is paraphrased as “why is this carrot rotten rather than fresh”, either of the two explanations above will do. However, if the question is rather posed as “why is this carrot rotten rather than this potato”, neither explanation suffices. The point here is that an explanation should not only explain why P, but also why not Q.

In order to be more precise, let us discuss the most prominent type of explanation within science, namely causal explanation (Salmon, 1984; Woodward, 2003). As a representative example, consider Woodward’s (2003; 2010) interventionist account of causation: “X causes Y if and only if there are background circumstances B such that if some (single) intervention that changes the value of X (and no other variable) were to occur in B, then Y or the probability distribution of Y would change” (Woodward, 2010, p. 290), where background circumstances are any circumstances concurrent with X and Y, irregardless of their causal relevance to Y, and an intervention is an idealized manipulation that only changes X, and causes a change in Y. A causal explanation therefore consists in demonstrating a pattern of dependence, such that changes in the explanans X are systematically associated with changes in the explanandum Y. Moreover, varying X should modulate Y in a fine-grained way, and the dependence between X and Y should not be altered by altering the background conditions. Importantly, "X has to be proportional to Y, meaning that X should not include irrelevant detail nor fail to include details that are relevant" (Woodward, 2010, p. 296f.), meaning that what we want to explain restricts the set of possible causes, in the contrastive manner described above. This is even more clearly stated in D. Lewis (1986), where causal explanation consists in citing a cause X for Y, that would not have caused another phenomenon Z, so X is the reason why Y and not Z.

Lastly, it is pertinent to discuss *mechanistic* explanation (e.g. Bechtel & Abrahamsen, 2005; Craver, 2007; Glennan, 2002, 2017; Machamer, Darden, & Craver, 2000), which is particularly relevant to the endeavours of cognitive science. According to Bechtel (2008), a mechanism is a "structure performing a function in virtue of its component parts" (p. 13). Mechanistic explanation consists in decomposing a mechanism into such component parts, and explaining it through their function and organization. An apt analogy is that of a mechanical watch: the watch is able to tell the time because of the organization of the individual cogs in the clockwork, and the different functions they perform. To explain a cognitive process, therefore, is to decompose the process into the various subprocesses, and their subprocesses, and so on, that jointly make the organism able to perform the process, and figure out which population of neurons is responsible for performing each of these subprocesses. This form of explanation thus consists in appealing to the lower-level parts, i.e. the components, of a mechanism to explain its higher level behavior.

### 3.2.2 How to Make ERPs Explanatory

Problematically, ERPs do not fit the mold on any of these representative notions of explanation, if we are trying to explain behavior. From a contrastive perspective, we might for example expect an ERN rather than an MMN when an unintended error is perceived, but it is not clear why we should expect a realization of unintended error rather than a perceived change in auditory stimulus when an ERN is detected. Consequently, it is the behavior explaining the ERP component, and not the other way around. A causal explanation does not fare any better. If we assume, e.g., that salience in the stimuli is caused by a P300, subtle changes in the P300 should induce subtle changes in the salience of the stimuli. Even setting aside the fact that subjects can detect salient stimuli without the P300 being visible in the EEG, at least on single trials, this assumption is absurd: since the P300 is defined in terms of the salience of the stimuli, there is no way to manipulate the P300 without manipulating the salience of the stimuli, thus violating the condition that no other variable in the background circumstances can be altered. However, if we assume that the salience of the stimuli causes the P300, we encounter no such problem, and it stands to reason that again it is the behavior explaining the neural response, which is not the intended direction of explanation.

From a mechanistic perspective, the main problem is that there has been a tendency in the cognitive sciences to characterize the operations underlying a behavior in terms of that behavior (Bechtel, 2005): in fact, a textbook explanation of an ERP component as "a scalp-recorded voltage change that reflects a specific neural or psychological process" (Kappenman & Luck, 2012, p. 4) reveals that ERP components are characterized exactly by the phenomena that they supposedly underlie. However, "the operations within a mechanism that enable it to perform its behaviors are [typically] of a different kind from those behaviors" (Bechtel, 2005, p. 320), and the main challenge for cognitive scientists is to conceptually conceive of what operations $\phi_1, ..., \phi_n$ would jointly consitute another operation $\psi$, without themselves being $\psi$. While the order and temporal signature of the various subcapacities can be discovered using ERPs, this information does not add anything to an explanation unless we know what a particular brain area is doing at a particular time, and we have concepts that can describe these activities.

If ERPs are not explanatory in and of themselves, their explanatory efficacy must lie elsewhere. Crucially, ERPs pick out particular temporal sequences of events, and is used to discover regularities between certain properties of a stimulus and amplitude changes in regions of the brain's electric field. Bogen points out that "the work generalizations do is epistemic rather than explanatory" (Bogen, 2005, p. 401), in that they can describe facts that warrant explanations, they can refine and make research questions more precise, they can restrict the space of possible explanations, and they can support inductive inferences required to theorize about causes or mechanisms. What is important conceptually, is that ERP components constitute an intermediate level that facilitates a relation between behavior and a cognitive mechanism in that you can rephrase your research question in the language of ERPs, a language that allows you to refer to lower-level processes that you otherwise would not have access to. For example, the fact that proportional quantifiers cannot be verified without the use of memory resources, generates a prediction that ERPs known to be sensitive to memory manipulations should be modulated by quantifier class. The use of memory resources would have to be inferred indirectly from behavioral measures such as accuracy and reaction time, whereas the use of ERPs allows one to measure the involvement of such systems directly.

Importantly, in order to avoid Poldrack's (2006) infamous *problem of reverse inference* – essentially, affirming the consequent – one must hypothesize what observations one should make on the basis of a theory, and subsequently attempt to falsify the theory by demonstrating that the expected outcome does not hold. This ties in with van Rooij and Baggio's (2020; 2021) criticisms of psychological theories in general. In two related papers, they argue that (1) psychology has been too focused on discovering effects, rather than on explaining capacities – as per the goals of cognitive neuroscience outlined above – and (2) that theories in psychology – presumably as a consequence – do not have rigorous hypotheses, but contain too many hidden assumptions. These considerations have prompted authors like Bird (2021) to examine the effects of weak hypotheses: if the probability that an hypothesis is true is low, the base error rate is much higher than rigorous, formal, and plausible hypotheses. Formalization is therefore a necessary, albeit not a sufficient, criterion for successful theory generation (van Rooij & Baggio, 2020). Formal rigor – e.g., as detailed in the previous chapter – is a prerequisite for making clear predictions, and if one is to utilize the intermediacy of ERPs, computational models of algorithms are essential for doing ERP research. With a clear theory about what constitutes a capacity – i.e. what a system is doing and how it is doing it – it becomes possible to formulate more fine-grained hypotheses that can utilize the ability to measure the recruitment of cognitive resources directly using EEG. For example, if some cognitive process is predicted to require more attention by the formal theory, and the evoked potential shows modulation of a component known to be modulated by attentional demands – e.g., the N2b (D'Arcy, Connolly, & Crocker, 2000; Wassenaar & Hagoort, 2007) – the formal theory explains the evoked potential, and the evoked potential corroborates the formal theory.

## 3.3 The Algorithmic Level

The idea that one needs to specify what a system is doing and how it does it is not new, however. In fact, the standard view in computational cognitive science, originates with Marr's (1982) three levels of analysis for information processing systems:

> **The computational level**: specifying a process as a function that takes a certain input to yield a specific output.
>
> **The algorithmic level**: detailing the stepwise procedures and subprocedures required to compute the function.
>
> **The implementational level**: describing how the algorithm is implemented in the physical medium of the brain to allow it to compute the function.

The computational and implementational levels are the subject matter for, in the present context, theoretical linguistics and neuroscience, respectively. In formal semantics, it is generally agreed upon that the function computed in semantic processing is taking a sentence as the input and outputting a truth value. By contrast, neuroscience describes the activities of neurons – i.e. receiving and transmitting electrical signals – and which populations of neurons are involved in different cognitive processes. How to translate the function from sentences to truth values into the reception and transmission of nerve impulses is not straightforward, however. This rendering presumably requires the gradual descent into subprocesses and their subprocesses discussed above, and requires considerable amounts of conceptual work. This is one of the reasons that the algorithmic level is vital in mediating between the computational and implementational levels. Another is the fact that the algorithmic level is constrained both by the computational level and the implementational level, in that the algorithm necessarily depends on the function to be computed, and also on the kinds of algorithms that can be implemented by a biological system such as the brain (Baggio et al., 2016, 2015; Embick & Poeppel, 2015; S. Lewis & Phillips, 2015).

Despite this fact, algorithmic aspects of semantic processing has not yet received sufficient attention (Baggio, 2018, 2020). This might be a consequence of meaning being notoriously illusive to formalization, which is required to construct algorithms, or that for many aspects of semantic processing, there are still disagreements about the computation being performed. Nevertheless, if we concede that knowing the truth value means knowing the meaning of a sentence – irregardless of one's opinions on the bijectionality of this statement – you can manipulate the truth value of sentences to observe the processing consequences of truth and falsity. This is in fact common practice, if not the default practice, in psycholinguistic research on semantics and pragmatics (Chemla & Singh, 2014; Katsos & Cummins, 2010; Noveck & Reboul, 2008; for a recent overview, see Cummins & Katsos, 2019). This gives you a well-defined procedure – verification – that, with the right implementation, can be operationalized in algorithmic terms.

At the very least, this is a computational problem. Analysis of the computational complexity of such problems, has been argued to constitute a kind of intermediate level between the computational and the algorithmic levels (Isaac, Szymanik, & Verbrugge, 2014). In fact, a growing body of literature (van Rooij & Baggio, 2020, 2021; van Rooij et al., 2019) has been advocating such analyses in psychological theorizing, generally. Complexity theory, as well as other tools from theoretical computer science, allows one to (1) demonstrate that certain computational problems are intractable, i.e. not computable by the brain, thereby falsifying claims about the computational level, (2) narrow down the possibility space for the algorithms that could be used to compute the computational level functions, and (3) prove that certain functions constitute more complex computational problems than other functions. The research in this thesis is focused on the latter two.

### 3.3.1 Experiments on the Algorithmic Level

Various complexity results have been used to study cognition, and I will give a few illustrative examples, beginning with the standard measures of computing time and computing space (see 2.2.1). Regarding direct measures, behavioral outcomes have been shown to be predicted by computational complexity for both social cognition (Szymanik, Meijering, & Verbrugge, 2013) and deductive reasoning (Gierasimczuk, van der Maas, & Raijmakers, 2013; Zhao, van de Pol, Raijmakers, & Szymanik, 2018) tasks. More indirectly, the fact that humans can approximate a near-optimal solution to the Euclidian travelling salesperson problem – which grows exponentially with the size of $n$ – in linear time (Dry, Lee, Vickers, & Hughes, 2006; Dry, Preiss, & Wagemans, 2012; Graham, Joshi, & Pizlo, 2000; van Rooij, Schactman, Kadlec, & Stege, 2006), has prompted researchers to search for *heuristics* or approximate solutions to such problems (Carruthers et al., 2018). In the case of the travelling salesperson problem, an hierarchical clustering algorithm produces a near-optimal solution because the precise order within a cluster does have a large impact on the overall effectiveness of the route (Graham et al., 2000). Importantly, this is not a different solution to the same problem, but, by definition, a distinct computational problem (van Rooij, Wright, & Wareham, 2012), and the tractability of these approximation problems should be studied in their own right (van Rooij & Wareham, 2012).

Also for language, the intractability of anaphoric reference resolution (Ristad, 1993) or the satisfiability of syllogistic reasoning with relative clauses (Pratt-Hartmann, 2004), have prompted scholars to investigate approximation algorithms for natural language as well (Pagin, 2012). More relevantly, complexity as measured by *expressive power*, impacts our language use, in that corpus frequency of linguistic constructions generally (Thorne, 2012), and quantifiers specifically (Szymanik & Thorne, 2017), is a function of computational complexity.

In fact, natural language quantifiers are an especially interesting case in this context. Since their semantic contribution is defined mathematically in generalized quantifier theory, their verification is a well-defined computational problem. We saw in the previous chapter that quantifiers fall into different classes depending on the automata associated with their verification. In particular, proportional quantifiers are associated with pushdown automata, with a memory component, whereas non-proportional quantifiers can be verified using a simple finite-state automaton, without such a memory component.

Two parallel series of studies have in fact demonstrated that this difference is manifested in real psychological, and even neural, effects. Szymanik and Zajenkowski have demonstrated that participants are less accurate and respond more slowly when verifying proportional quantifiers, compared to non-proportional (Szymanik & Zajenkowski, 2010a). These effects are modulated by memory load (Szymanik & Zajenkowski, 2010b, 2011), and are correlated with participants' working memory capacity (Zajenkowski et al., 2011; Zajenkowski & Szymanik, 2013; Zajenkowski et al., 2014). McMillan and colleagues have shown that these differences are also reflected in neural activity. Using fMRI, they demonstrated a larger BOLD response in (pre)frontal areas associated with working memory and executive function, notably the dorsolateral prefrontal cortex, during the verification of proportional compared to non-proportional quantifiers (McMillan et al., 2005; Olm et al., 2014). Patients with focal neurodegenerative disorders affecting these regions, are also impaired with proportional quantifiers only, and the degree of atrophy is correlated with performance in a verification task (McMillan et al., 2006; Morgan et al., 2011). Interestingly, this resonates with fMRI studies from the mathematical cognition literature (Jacob & Nieder, 2009; Mock et al., 2019, 2018), where bilateral frontal activation is associated with the processing of proportions in adaptation and magnitude comparison paradigms, irrespective of the mode of presentation – mathematically or verbally.

Taken together, these findings suggest that algorithmic aspects of semantic processing – as well as other cognitive processes – manifest as cognitive costs, using a variety of psycholinguistic measures. Regardless of the specific algorithms people use, the minimal complexity of the computational problem can be viewed as a lower bound, and it is this fact that makes complexity analyses useful in psycholinguistics.

## 3.4 Methodological Limitations and Precise Implementation of the Research Questions

Since the complexity theory approach to EEG data has not been taken extensively in the past, it is beneficial to implement a rather conservative paradigm. This is because we want to as much as possible isolate only the impact of the algorithmic analysis, and not have uncertainties about the influence of novelties of the paradigm, the stimuli or the methods of statistical analysis or preprocessing. Since ERPs are, arguably, the most well-understood approach to EEG data (Rommers & Federmeier, 2018), at least in sentence processing contexts, it was therefore decided to conduct an ERP study of picture-sentence verification. The choice of the task was motivated by two main considerations: (1) the fact that all previous examinations of the impact of complexity have implemented such a task (e.g. McMillan et al., 2005; Olm et al., 2014; Szymanik & Zajenkowski, 2010a; Zajenkowski et al., 2014), and (2) the fact that the mathematical proofs are about verification procedures, using objects as sequential input.

EEG has some obvious disadvantages when it comes to picture-sentence verification: the detectable brain signals at the level of the scalp is very weak, much weaker than those caused by for example muscle movement. This means that any form of movement, and in particular eye-movement,[1] distorts the EEG signal and obscures the concurrent brain activity (Luck, 2014; Plöchl, Ossandón, & König, 2012). While numerous techniques have been suggested to subtract the eye-movement artefact, or at least minimize its impact on statistical analysis, traditionally such trials are discarded. So if participants were, e.g., scanning a picture while listening to a sentence, the amount of eye-movement could render the data effectively useless. Because of this, pictures were presented before each sentence, and only the EEG data recorded during reading was subjected to analysis. However, this increases the difficulty of the task considerably, and consequently restricted the potential complexity of the images. The pictures therefore consisted of clusters of red and yellow circles and triangles, the minimal number of objects (2 × 2) that allowed us to control when participants could know the truth value of the sentences. The number of objects in a cluster were also relatively low (2-5). Furthermore, a block design was implemented, where one picture was shown before every trial in a block. This was (1) to ensure that participants could remember the picture, since this is necessary for performing the task, and (2) to minimize differences in encoding or recall of the picture between trials. While one might, perhaps rightfully, worry that all quantifier classes require some form of memory to be verified with this paradigm, the automata theory demonstrates that *proportional* quantifiers requires additional memory resources to maintain and compare two sets of objects in memory, which the other classes do not. It is therefore predicted that proportional quantifiers should recruit additional memory resources compared to their non-proportional counterparts. If anything, the stable baseline of this set up plausibly relates any observed differences to the experimental manipulation, rather than differences in encoding or retrieval.

For similar reasons, the standard way of presenting sentences visually in ERP paradigms is through *serial visual presentaion* (SVP), where a sentence is presented in chunks at a fixed rate. Since this is the perceived standard, SVP was used in all experiments conducted for this thesis. However, it should be noted that the mode of presentation, SVP or (semi)natural aural presentation, has been shown to impact the processing of quantified sentences (Freunberger & Nieuwland, 2016). Nevertheless, SVP has an additional advantage in the present context. Since quantifier expressions differ in length (compare e.g. ‘all’ to ‘more than half’), this difference needs to be controlled for. In an SVP set up, we could easily present the quantifier as one chunk,

[1] For a discussion of the constant electrical potential between the cornea and retina, see e.g. Luck (2014).

regardless of its length, thus ensuring that processing of the remaining words of the sentence, presented individually, would be identical between quantifiers.

Turning to the sentences specifically, these were simple copular sentences where a color (red or yellow) were predicated of a definite quantified noun phrase with the shape noun (circle or triangle) as the head. This choice was necessitated by the nature of the pictures. For a discussion of the specific quantifiers used, the reader is referred to the respective papers, and for general considerations about quantifiers in Norwegian, see 3.5 below.

Since the ERPs of the distinct quantifier classes are in uncharted territory, it was useful to design the experiments so that known ERP components were manipulated as well. In sentence processing, the N400 and the P600 components are the most studied (Swaab et al., 2012). The P600 – a positive shift in the ERP waveform that is largest around 600 msec – has been associated with increased difficulty in integrating the incoming word into the wider context (Brouwer & Hoeks, 2013; Delogu, Brouwer, & Crocker, 2019), decision complexity (Sassenhagen, Schlesewsky, & Bornkessel-Schlesewsky, 2014), as well as with composition (Baggio, 2021; Fritz & Baggio, 2020, 2021). The functional significance of the N400 – a negative deflection in the ERP peaking around 400 msec – is a matter of debate. Specifically, scholars disagree about whether a larger N400 reflects increased difficulty in retrieval (Brouwer, Fitz, & Hoeks, 2012; Kutas & Federmeier, 2011; Lau, Almeida, Hines, & David, 2009; Lau, Phillips, & Poeppel, 2008), integration (Brown & Hagoort, 1993; Hagoort, Hald, Bastiaansen, & Petersson, 2004), or both (Baggio & Hagoort, 2011; Nieuwland et al., 2020). Regardless of the specific cognitive process it underlies, however, its amplitude is predominantly modulated by semantic association (Aurnhammer, Delogu, Schulz, Brouwer, & Crocker, 2021; Kutas, 1993), frequency (Laszlo & Federmeier, 2014; Rugg, 1990; Van Petten, 1990, 2014), and contextual predictability (Delogu, Crocker, & Drenhaus, 2017; Nieuwland, Ditman, & Kuperberg, 2010; Nieuwland & Kuperberg, 2008; Nieuwland & van Berkum, 2006). Since we are dealing with the semantics of quantifiers, and the N400 is traditionally associated with semantic violations, the N400 would arguably be the most likely candidate to manipulate.

In a picture-sentence verification task, false sentences are incongruent with the preceding visual context, and should thus elicit a larger N400 than true sentences (Augurzky, Bott, Sternefeld, & Ulrich, 2017; Knoeferle, Urbach, & Kutas, 2011; Nieuwland & Martin, 2012). As a sanity check, we therefore manipulated truth value, so that half the sentences were true and half the sentences were false. If an N400 was observed for false versus true sentences, it could be inferred that participants were performing the task correctly. Since the stimuli at the sentence final adjective – the earliest position where the truth value could be unambiguously determined – were otherwise identical, no other factors known to influence the N400, such as frequency and semantic association, could have an effect.

The schema of the experimental design was thus a picture-sentence verification task, where a picture was presented before a copular sentence, which was presented in chunks. The experiments had a blocked design, were the picture remained the same within a block. The independent variables were Quantifier Class and Truth Value. Aside from the quantifier, all sentence were syntactosemantically identical. This allowed us to (1) know whether participants were verifying the sentences correctly without worrying about movement artefacts, (2) isolate the differences stemming from the quantifier class manipulation, and (3) investigate at what stages of processing potential differences between verification algorithms manifest themselves. Importantly, the nature of the difference, which is predicted to be related to memory, could not be inferred from this bare experiment, so additional measures – i.e. a direct manipulation of memory load – had to be implemented to answer this research question. For details about the concrete implementation of the experiments – e.g. stimuli, EEG recording etc. – as well as for modifications of the base design, the reader is referred to the respective papers, 1 and 2.

## 3.5 A Note on Norwegian Quantifiers

Before summarizing the chapter, a brief interlude on some aspects of quantification in Norwegian is required. Generally, Norwegian does not differ in any significant way from other languages when it comes to quantification. It generally patterns with its Scandinavian siblings and its Germanic cousins, both in terms of syntax and semantics of various quantifier expressions. In particular, all quantifier meanings discussed thus far have Norwegian equivalents, and there are quantifiers from all the four quantifier classes.

However, parity quantifiers are somewhat marked in Norwegian. Both 'et like antall' (*en even number*) and 'et odde antall' (*an odd number*) are attested, but rather infrequent (see table 3.1).[2] While this is not necessarily a problem, they cannot take a definite complement, which was necessary to be matched with the other conditions in the experiments we conducted, as some of the quantifiers require definiteness to be referential. One point of contention between the two parallel research communities that have investigated complexity effects in the verification of quantifiers previously, is precisely the status of parity quantifiers: whether they should be grouped with proportional or non-proportional quantifiers (see Szymanik,

[2] It might also be noteworthy that for the corpora where source data is available, these expressions overwhelmingly come from math and science related domains, as well as chess problems. This is an indication that they might only be available in specific registers, and that they are not commonly used in vernacular language.

| Quantifier | NoWaC | LB | HaBiT BM | HaBiT NN |
|---|---|---|---|---|
| Et like antall | 5 | 1 | 89 | 1 |
| Et odde antall | 4 | 5 | 41 | 1 |
| Corpus size | 700 | 100 | 1180 | 55 |

Table 3.1: Corpus frequency of parity quantifiers from the Norwegian Web As Corpus (NoWaC), Leksikografisk bokmålskorpus (LB), Harvesting big text data for under-resourced languages (HaBiT) bokmål (BM) and nynorsk (NN) corpora. Corpus size is reported in millions of tokens.

2007; Szymanik & Zajenkowski, 2009; Troiani, Peelle, McMillan, Clark, & Grossman, 2009b). The automata theory predicts them to pair with other non-proportional quantifiers, and it would therefore be advantageous to be able to demonstrate this using brain based methods, as this has not been done before. Since this was not possible in our experiments, future experiments should put this hypothesis to the test in languages where such expressions are less marked.

Another quirk of Norwegian, as well as the other Scandinavian languages, is that proportional quantifiers – or *quantity adjectives* as they are referred to as in the literature on the semantics of degree – has the opposite definiteness pattern compared to English. In English, bare 'most' has the proportional reading, i.e. *more than half*, whereas definite 'the most' has a relative meaning, e.g. John read more books than any other contextually salient individual in a sentence like 'John read the most books'. For Norwegian, 'de fleste' has the proportional reading, whereas 'flest' has a relative reading. For a more thorough review, see Coppock (2019).

Hackl (2009) famously argued that 'fewest' does not have a proportional reading, since, contarary to the biggest subset 'most', there is no unique smallest subset: any singleton set has the same cardinality. 'De færreste' is nevertheless commonly used in Norwegian to express a small minority, and could be paraphrased as *very few, possibly no*. However, this meaning is not proportional and could therefore not be used as the opposite of 'de fleste'. In order to control for the polarity of the proportional quantifiers, we chose to include 'færrest av'. As can be inferred from the preceding paragraph on definiteness, this gives a relative reading of the quantifier, in contrast to the proportional quantifier 'de fleste'. The justification for this decision was that since there were only two relevant subsets (e.g. yellow and red triangles), the contextually salient relatively smallest subset and less than half were denotationally equivalent. This ensured complementarity of the proportional quantifiers – i.e. that one was true in contexts where the other was false – but the discrepancy between the two quantifier readings, prompted us to change 'de fleste' to 'flest av' for the second study, reported in paper 2. It is important to note that 'færrest av' is much more marked and less frequent than 'de fleste' or 'flest av'. However, this presumably stems from a bias against downward monotone quantifiers generally[3] (Szymanik & Thorne, 2017), and could not be avoided if we wanted to match proportional quantifiers for polarity. Because the quantifiers were easily interpretable in context – as is evidenced also by the high accuracies for negative proportionals in all experiments – symmetry in the design was allowed to prevail.

## 3.6 Chapter Summary

This chapter has dealt with philosophical considerations about explanation in cognitive science, as well as mundane facts about Norwegian quantifiers and practical considerations of the limitations of experimental implementation. I have attempted to outline the nature of these opposing methodological considerations and to describe how they interact, in order to motivate the specific way that the experiments in this thesis were designed.

After describing the two main ways of studying human neural activity and justifying the choice of event-related potential (ERP) analysis of electroencephalography (EEG) as the method for the experiments conducted for this thesis, I presented some foundational problems with this method. Importantly, I argued that ERPs are explananda rather than explanantia, and that their contribution to our understanding of cognitive processes are dependent on the theory that explains them. I further stressed the need for algorithmic and complexity theoretic analysis in explanations in Marrian cognitive science, and highlighted ways in which such analyses have been used in the past. These considerations were used to argue for the experimental design. Participants performed a picture-sentence verification task with Truth Value and Quantifier Class as independent variables, where the picture was presented before the quantified sentence. Finally, some peculiarities of the Norwegian quantifier system were used to explain certain design choices.

[3] The same is true for 'more/less than half', 'the majority/minority of' as well.

# Chapter 4
# Summary of Papers

## Paper 1:
### Computational Complexity Explains Neural Differences in Quantifier Verification

Paper 1 consists of two experiments, designed to (1) observe differences between the quantifier classes, and (2) to assess task effects of verification.

Participants ($N = 24$ in both experiments after artifact rejection) performed a picture-sentence verification task, as described in chapter 3, in experiment 1 and saw the same stimuli in experiment 2, but were not required to verify the sentences. Instead, they were asked simple comprehension questions about the stimuli. The design for both experiments was $3 \times 2$, with Quantifier Class (3 levels: Aristotelian, Numerical, Proportional) and Truth Value (2 levels: True, False) as independent variables. 1000 msec epochs, including a 200 ms baseline, extracted from the noun completing the subject noun phrase and from the predicate adjective, were subjected to pairwise comparisons using cluster based permutation statistics (Maris & Oostenveld, 2007).

In experiment 1, the False-True comparison yielded sentence final negativities in the 2-500 msec time window for all Quantifier classes, both individually and overall. Interestingly, the effect of Truth Value was modulated by Quantifier Class, such that the negative effect was largest for Aristotelian quantifiers and smallest for Proportional, with Numerical quantifiers falling in between. The negative effect for Proportional quantifiers was followed by a positive cluster in the P600 time-window. Since the Truth Value effect presupposes that participants know the truth value, and therefore by definition that a verification procedure has been completed, we expect potential verificational differences to occur earlier in the sentence. When comparing the Quantifier classes at the noun, we found a late positive cluster for Proportional quantifiers relative to the other two, both individually and collapsed.

In experiment 2, by contrast, the negative effect of the False-True comparison was only significant overall and for Aristotelian and Numerical quantifiers. The positive effect of Proportional quantifiers at the noun also disappeared completely.

These results indicate that the complexity differences between Quantifier classes manifest as real electrophysiological effects during verification, with additional effects on subsequent processing. Importantly, when verification is not required to perform the task, participants are able to track the truth value of Aristotelian and Numerical quantifiers only, presumably because Proportional quantifiers are too complex to verify tacitly. The modulation of the Truth Value effect in experiment 1, is further evidence of such downstream consequences of the complexity of verification. The effects not directly pertaining to verifiation is beyond the predictive scope of the automata theory, and any interpretation of these are therefore speculative. However, the automata theory predicts the difference between Proportional and Non-Proportional quantifiers to be the result of a memory component, and the fact that late positivities have been associated with task-relevant recollection memory (e.g. Rugg & Curran, 2007), therefore consitutes preliminary evidence for the involvement of memory systems in the verification of Proportional quantifiers.

# Paper 2:

## The Interplay of Computational Complexity and Memory Load in Quantifier Verification

The aim of the experiment in paper 2 was to ascertain whether the differences between proportional and non-proportional quantifiers were related to memory. To this end, we modulated memory load during verification and collected measures of working memory from the participants.

Participants ($N = 48$ after artefact rejection) completed a working memory battery, before performing the same sentence-verification task as in the previous study. While they were performing the verification task, they had to remember 2 or 4 digits presented at the beginning of each trial, and judge whether it matched a second string of digits at the end of the trial. The design was thus $2 \times 2 \times 2$, with Quantifier Class (Proportional, Non-Proportional), Digit Load (2, 4 Digits) and Truth Value (True, False) as independent variables. Data analysis was as close to identical as that in study 1.

The sentence final effects were replicated, with False sentences being more negative than True in the same time-window, with a smaller difference for Proportional than for Non-Proportional, and a later positivity for Proportional quantifiers. Additionally, there were main effects of Quantifier Class and Digit Load, such that Proportional Quantifiers were more negative than Non-Proportional and 4 Digits were more positive than 2. The effects at the noun were of a completely different nature, however. Instead of a late positivity, the Proportional-Non-Proportional comparison yielded an early left hemispheric negativity, more consistent with sustained left anterior negativities (LAN) or sustained anterior negativities (SAN) (e.g. Baggio, van Lambalgen, & Hagoort, 2008; Fiebach, Schlesewsky, & Friederici, 2001; van Berkum, Brown, Hagoort, & Zwitserlood, 2003; Vos, Gunter, Kolk, & Mulder, 2001), associated with increased working memory demands. This effect was modulated by Digit Load, such that it was only significant overall and for 4 Digits, but a general linear model of cluster amplitude revealed that the interaction was not significant. Participants' working memory scores did not predict individual cluster amplitude or effect size, possibly as a result of low variability in the scores.

The results provide further indicative evidence of the effects of computational complexity on verification and on subsequent processing. Moreover, the fact that Digit Load affects the Quantifier Classes differently, in particular during verification, more firmly suggests that the difference is related to memory, despite the lack of a significant interaction effect and correlations with individual working memory capacity. However, the blatant difference in effect type – LAN/SAN, rather than LPC – suggests that the specific memory systems employed by the brain may differ depending on the task. Despite this fact, we argue that the constraints on human quantifier verification is of the same nature as the constraints on abstract machines.

## Paper 3:

### Neural Algorithms of Natural Language Quantification: A review of the experimental literature

In paper 3, I attempt to demonstrate that the algorithms of the quantifier classes are deducible from the extant experimental literature on quantifier verification in the brain, given the assumption that people should use the simplest possible algorithm to solve a computational problem (Anderson, 1990; Szymanik, 2016). Since few studies have compared quantifier classes directly, I survey the studies looking at only one quantifier class as well.

Aristotelian quantifiers are predicted to primarily rely on attentional mechanisms required to detect examples or counterexamples. While there is some conflicting evidence regarding the recruitment of magnitude processing in the intraparietal sulcus (IPS) – which is strictly not required for Aristotelian quantifiers – experiments that explictly require verification (Morgan et al., 2011; Olm et al., 2014) do in fact observe (pre)frontal activation associated with attention and detection. ERP data consistent with these findings reveal that such mechanisms are activited only when participants are able to unambigously determine the truth value of the sentence, where a biphasic pattern of an early negativity and subsequent positivity observed with false sentences is constituted by the detection of a mismatch and a subsequent increase in attention to this counterexample in order to judge the sentence as false, in line with the algorithm for Aristotelian quantifiers (Augurzky et al., 2017). The inter- and intraindvidual differences resulting from scalar implicature reading of 'some' (*some and not all*), are also discussed, as are the complexity consequences of multiply quantified sentences.

Numerical quantifiers rely on a counting algorithm for their verification, and they are therefore predicted to involve magnitude processing in the IPS, both by the automata theory and the *Triple code model* of mathematical cognition (Dehaene, 1992; Dehaene, Piazza, Pinel, & Cohen, 2003). This prediction is showed to be supported by the available empirical evidence. Szymanik's (2016) prediction that the increase in number of states for higher numbers should lead to an increase in complexity is discussed, but is set aside for future research due to insufficient data. The availble data also suggests a preference for an *exact* – as opposed to an *at least* – reading of numerals for most people.

The available neurolinguistic literature concerning parity quantifiers, which, like Aristotelian quantifiers, should primarily consist in the deployment of executive resources such as cognitive control (Szymanik & Zajenkowski, 2009), is sparse. In particular, such quantifiers have never been separated out in the analysis in verification experiments. While the mathematical cognition literature has revealed differences between parity and magnitude estimation, and between judging the parity of dot arrays rather than numerals, future research is needed to determine whether human subjects are using the minimally complex algorithm to determine parity in a verification task.

Proportional quantifiers are the only quantifiers that need the additional memory resources of pushdown automata (PDAs) to be verified. The prediction is therefore that additional memory resources should be recruited during proportional quantifier verification. Both fMRI and patient studies point to the involvement of (pre)frontal areas as well as the IPS during proportional quantifier verification, indicating that people are estimating the size of the sets and comparing them by means of working memory. However, EEG studies have shown that complicating the task, gives rise to different evoked responses to proportional quantifiers, thereby obscuring the linking hypothesis between the theoretical stack memory and memory resources in the human brain. Finally, semantic and pragmatic properties irrelevant to verificational complexity – such as negative polarity – are shown to impact the incrementality and processing difficulty associated with proportional quantifiers.

The paper concludes with a discussion on what measures of complexity are relevant to quantifier verification, and stresses the importance of embedding formal results in plausible models of human cognitive processing.

# Chapter 5
# A Synthesis of the Results

Having summarized the experimental work, it remains to elaborate on what conclusions can be drawn. This chapter is devoted to answering the research questions presented in chapter 1 and repeated below, in so far as it is possible.

(1) Do the differences in the computational complexity of verification algorithms for proportional and non-proportional quantifiers manifest in distinct brain responses?

(2) If so, are these brain responses related to memory, as predicted by the automata theory?

(3) At what point(s) during sentence processing do such differences occur?

(4) What model of sentence processing best explains such patterns?

Recall that in order to answer these research questions, the available literature on quantifier processing using neural measures was reviewed (Paper 3), and three experiments were conducted (Papers 1 and 2). These experiments utilized event related potential (ERP) analysis of electroencephalography (EEG) data, and had a picture-sentence verification task at its core. The need for explicit verification (Paper 1) and the modulation of memory load by the addition of a digit matching task to the experimental paradigm (Paper 2) was manipulated, in order to ascertain whether the observed effects were related to verification and to memory, respectively.

The differences afforded by algorithmic distinctions is presented in 5.1, and 5.2 is devoted to their relation to the employment of memory resources. The model of sentence processing suggested by the data and the time-course of verificational differences is the subject of 5.3. 5.4 summarizes the discussion by way of explicit answers to the research questions. A general outlook and directions for future research is reserved for Chapter 6.

## 5.1 Distinct Neural Patterns of Quantifier Automata

As is evident from the experimental work presented herein, as well as plausibly inferred from the previous literature, the minimal complexity of the verification algorithm is associated with distinct brain responses. Interestingly, the quantifier classes do not only diverge during verification proper, but impacts subsequent processing related to prediction and decision processes as well. However, since such differences are not strictly predicted by the automata theory, they can only be explained in the context of a model of sentence processing. Hence they are left to section 5.3.2 below.

The theory predicts proportional, but not non-proportional, quantifiers to trigger known brain signals of memory. This is indeed what we find. As discussed in the review in paper 3, it can be inferred from the patient studies that quantifier comprehension can be selectively impaired, depending on the computational resources required to compute them (McMillan et al., 2006; Morgan et al., 2011; Troiani et al., 2009a). Behavioral variant fronto-temporal dementia (bvFTD) patients, whose symptoms include reduced executive functioning, display poorer behavioral results for Aristotelian and proportional quantifiers, while corticobasal degeneration (CBD) patients, who suffer from acalculia, were impaired in the verification of numerical and proportional quantifiers. The verification scores are also mostly correlated with relevant behavioral measures of cognitive function – e.g. digit span, stroop color naming etc. – as well as with atrophy in specific regions of the cortex: (pre)frontal areas for the bvFTD and temporo-parietal areas for the CBD patients. BOLD fMRI results of healthy participants (McMillan et al., 2005; Olm et al., 2014) also implicate these areas in processing, such that Aristotelian quantifiers recruit (pre)frontal areas, numerical quantifiers recruit parietal areas, and proportional quantifiers recruit both areas equally. A similar activation pattern to that of proportional quantifiers is found for proportions in the mathematical cognition literature as well (Jacob & Nieder, 2009; Mock et al., 2019, 2018). (Pre)frontal areas are typically associated with working memory, attention and other executive functions (Aron, Robbins, & Poldrack, 2014; Badre & Wagner, 2004; Brunoni & Vanderhasselt, 2014), whereas the intraparietal sulcus (IPS) is hypothesized to be the center of magnitude processing (Dehaene, 2011; Nieder & Dehaene, 2009; Skagenholt, Träff, Västfjäll, & Skagerlund, 2018).

This pattern is explained by the automata theory. Because Aristotelian quantifier verification only relies on detecting (counter)examples, counting or other kinds of magnitude processing is not necessary to perform the task. For numerical quantifier verification, by contrast, only counting is necessary, and the extent of working memory usage limits itself to remembering the last number counted.[1] Lastly, the algorithm for proportional quantifiers involve both magnitude processing – comparing the As that are B and the As that are not B – and, crucially, memory – keeping track of both subsets of A. It is therefore exactly as predicted that Aristotelian and numerical quantifiers are processed predominantly in the (pre)frontal and parietal cortices, respectively, whereas proportional quantifiers recruit both cortical areas for their verification.[2]

Analogs of these differences are also found in evoked potentials. During simple picture-sentence verification, proportional quantifiers elicit a late positivity compared to non-proportional quantifiers upon the completion of the subject noun phrase (Bremnes, Szymanik, & Baggio, 2022; De Santo et al., 2019). Such positivities have been associated with recollection memory, where the so-called *late positive complex* (LPC) is a component that appears when recollecting the details of a stimulus is task relevant (Hubbard, Rommers, Jacobs, & Federmeier, 2019; Ratcliff, Sederberg, Smith, & Childers, 2016; Rugg & Curran, 2007; Rugg et al., 1998; Yang et al., 2019). This effect is similar in distribution to the P600, which has also been linked to episodic memory (O'Rourke & Van Petten, 2011; Van Petten & Luka, 2012), albeit in single word contexts, and not during sentence processing. In light of the automata theory, this would seem to indicate that participants recollect more details of the picture – e.g. both red and non-red circles – when verifying proportional quantifiers, than when verifying non-proportional quantifiers – e.g. only red circles. However, since De Santo et al. (2019) also observe a similar positivity in their study even though participants are looking at the picture while hearing the sentence, such an interpretation is not unproblematic. One worry is that this P600-like positivity indexes generic processing costs (Brouwer & Hoeks, 2013; Delogu et al., 2019), rather than recollection memory. But as described in chapter 3, this is an unwarranted reverse inference. Since ERPs are explananda, and not explanantia, it is the theory that does the explaining. The memory component predicted by the automata theory is a more specific explanation than the generic processing cost explanation (Woodward, 2010), and is consequently superior, i.e. deployment of memory is a specific processing cost. Because the theory is not falsified by the

[1] It has been hypothesized that the complexity of the numerical verification algorithm should increase with higher numbers (Szymanik, 2016): keeping track of the number of As that are B when counting for a long time could potentially trigger some form of working memory, even though it is not necessary (but see De Santo and Drury (2020) and Shikhare, Heim, Klein, Huber, and Willmes (2015) for a discussion about how this might be confounded with the magnitude comparison ratio). However, the numerical quantifiers used in the aforementioned experiments have all been small numbers, e.g. 'three', 'five' etc. Since the hypothesis has not been put to the test, it is not possible to ascertain whether this is in fact the case.

[2] See also Fitch (2014) for a proposal that the stack memory of pushdown automata is instantiated by mechanisms in the inferior frontal gyrus.

extant experimental data, abductive inference favors the memory interpretation. Opponents of this interpretation, should therefore provide a specific processing cost that is not related to memory.

One candidate is integration into the wider linguistic context (e.g. Brouwer & Hoeks, 2013). However, this presumably falls into the category of syntacto-semantic composition effects, since the noun is identical across conditions and the only difference between the noun phrases is the preceding quantifier: integration into the linguistic context is equivalent to noun phrase composition in the experiments under discussion. Such effects were not found in the comprehension study in paper 1 (Bremnes et al., 2022), and consequently this interpretation arguably does not fit the data equally well. One might argue that participants are not actually composing sentences, but only looking for the specific linguistic and pictorial material required to answer the comprehension questions. But this leaves the effects of the other two quantifier classes unexplained. Participants display similar ERP patterns throughout the sentence for Aristotelian and Numerical quantifiers in both experiments, and if the lack of composition is selective to proportional quantifiers, this necessitates an explanation of why these quantifiers are different. The obvious answer is provided by the automata theory, thus making the argument circular.

Interestingly, the evoked potentials of verification differ with the addition of the digit matching task in paper 2. At the same place in the sentence, i.e. at the onset of the noun, proportional quantifiers trigger a left hemispheric negativity in the 250-500 msec time-window relative to non-proportional quantifiers, instead of the positivity in the P6 time-window. The spatiotemporal distribution of this effect is consistent with sustained *left anterior negativities* (LAN) (Fiebach et al., 2001; King & Kutas, 1995; Kluender & Kutas, 1995; Vos et al., 2001) or *sustained anterior negativities* (SAN) (Baggio et al., 2008; Müller, King, & Kutas, 1997; Münte, Schiltz, & Kutas, 1998; van Berkum, Brown, & Hagoort, 1999; van Berkum et al., 2003), both of which have been associated with working memory during sentence processing. This suggests that the specific memory systems recruited to verify a sentence when the verification procedure requires memory, differs depending on the nature of the task. For some tasks, like digit matching, simple recollection might not be sufficient, and thus working memory resources are engaged. When working memory systems are already recruited, like in the digit matching task, these systems might be utilized by the verification algorithm as well, on the assumption that it is better to deploy more of the same cognitive resources rather than engaging another system. When these systems are not already recruited, like in the simple verification task in paper 1, other memory systems might be deemed more apt. Another option is that both systems are recruited during verification with digit matching, but that the late positivity associated with recollection memory is obscured by the earlier negativity.

Which alternative turns out to be correct is not predicted by the formal results, and therefore remains an empirical question that is subject to additional assumptions.

This result is in a way not surprising. As discussed in the introduction, mathematically specified theories are independent of implementational detail (van Rooij & Baggio, 2021). The formal theory serves to isolate indispensible properties of a pattern, that any computational device must be able to detect in order to perform a computational problem (De Santo & Rawski, 2022): what is important is that the algorithms share these essential properties, not the specific details of their implementation. Nevertheless, it might be useful to disentangle the different notions of memory in operation.

## 5.2 Complexity and Memory

There are at least three notions of memory at play in the experimental studies in this thesis:

(1) The abstract semimetaphorical stack memory of the pushdown automaton

(2) The memory system that implements this stack in human brains

(3) The neural memory systems involved in recalling the contents of the picture preceding each experimental trial

Obviously, these three notions are interrelated. In particular, the relationship between (1) and (2) is just a linking hypothesis; (2) is, by definition, the memory component that instantiates the PDA stack memory. The relation between (2) and (3) is slightly more complicated. Whether (2) and (3) represent distinct memory systems, or whether they are in fact the same, is an empirical question. The results from paper 1 (Bremnes et al., 2022) seem to suggest that they are the same thing, since the memory component that seems to be modulated, the LPC, has previously been related to task relevant recollection (Rugg & Curran, 2007). However, the early negativity observed in paper 2, casts doubt on this simple relation. Since the LAN and SAN components are associated with working memory and not with recollection, this result seems to indicate that, at least sometimes, the implementation of the stack memory and the memory systems underlying recall of the picture can come apart.

This is potentially a problem, since the prediction of the automata theory is that (1), therefore (2), but what we might actually be observing, at least in paper 1, has more to do with (3). The crux of answering the research question of whether the differences are related to memory consequently lies in specifying the precise relation between (2) and (3).

While the evidence presented in the thesis, along with the extant experimental literature, is not conclusive, there are some important insights to be had. As mentioned in 5.1 above, De Santo et al. (2019) observed similar effects to those of paper 1 (Bremnes et al., 2022), as well as other picture-sentence verification studies (Augurzky et al., 2017; Augurzky, Schlotterbeck, & Ulrich, 2020), even when participants do not have to recall the picture, thereby casting doubt on this component's relation to recollection. There are other positive components, notably the *positive slow wave* (PSW), in the more explicitly working memory related literature (Kusak, Grune, Hagendorf, & Metz, 2000; Lefebvre, Marchand, Eskes, & Connolly, 2005; Marchand, Lefebvre, & Connolly, 2006; McEvoy, Smith, & Gervins, 1998; Pelosi, Hayward, & Blumhardt, 1995, 1998; Pelosi et al., 1992; Ruchkin, Johnson, Grafman, Canoune, & Ritter, 1992) with a similar spatiotemporal distribution, argued to index retrieval from short-term memory (García-Larrea & Cézanne-Bert, 1998), that could be a plausible interpretation of the late positive effect from paper 1. The involvement of brain regions associated with working memory during picture-sentence verification without recall (McMillan et al., 2005, 2006; Morgan et al., 2011; Olm et al., 2014) could support such an interpretation.

If this is correct, the relation between (2) and (3) is one of no identity and no overlap. The effects we see are solely related to working memory differences, and the effect of recall is not observed. This would also make sense since the recall of the picture, by design, is constant between the quantifier classes; the same images are shown before all quantifier types, and the only difference between them is the complexity of the verification task. In order to corroborate this hypothesis, a study would have to be conducted where the complexity of the recall is not fixed. The hypothesis would be corroborated if the difference between the quantifier classes was not modulated by recall complexity. If, on the other hand, we observe an interaction effect between the complexity of recall and quantifier class, this hypothesis would be falsified.

However, this leaves us without an explanation for the diametrically different effect that was observed in paper 2. It is unclear why different memory systems should be recruited under memory load, if the memory system recruited in paper 1 and in De Santo et al. (2019) is also a working memory system. Above, I suggested that working memory resources were recruited when systems of recollection memory were insufficient to perform the task. If this is not the case, some other explanation is required. One possible such is that the addition of the memory load recruits verbal or central executive working memory, whereas the simple picture-sentence verification task only relies on visual working memory (for discussion see e.g. Baddeley, 2012). A problem with this interpretation is that visual working memory is associated with negative effects (Axel & Müller, 1996; Rösler, Heil, & Röder, 1997; Ruchkin, Johnson, Canoune, & Ritter, 1990; Ruchkin et al., 1992; Vogel & Machizawa, 2004), and the component manipulated in the proposed visual working memory situation is positive. A solution along these lines is therefore not straightforward, and is, arguably, not tenable without additional assumptions and experimental evidence.

Hence, one major remaining open question left by the experimental results presented herein, is precisely the nature of the memory component and the relation between (2) and (3) above. Future research should be designed to better understand which contexts facilitate the late positivity observed in paper 1, and in which contexts an early negativity, like the one observed in paper 2, is observed instead.

## 5.3 Verification and Sentence-Processing

Since we have explored sentence processing both without an explicit verification task as well as with explicit verification – with and without memory load – we are able to disentangle the ways in which verificational properties of quantifiers interact with participants' processing of declarative sentences. There are a couple of ways in which the present results lend themselves to a precise model of sentence processing, most important of which is the fact that participants are sensitive to the truth value of the sentence they are reading. I will therefore deal with this issue first, in 5.3.1, before going on to discuss how this model is affected by the complexity of the verification algorithm in 5.3.2.

### 5.3.1 A Model of Sentence Processing

It is evident that participants keep track of the truth value of the sentences they are reading, since true and false completions of a sentence elicit distinct electrophysiological responses: false sentences are characterized by a large negative deflection relative to true after around 250 msec following the onset of the sentence final adjective. The model of sentence processing most consistent with the data, is therefore one in which participants are building a true model of the sentence in the context of a preceding picture (Baggio, 2018; Clark, 1976; Clark & Chase, 1972, 1974; Johnson-Laird, 1983; Just, 1974; Just & Carpenter, 1971; van Lambalgen & Hamm, 2005; Zwaan & Radvansky, 1998). In other words, the participants expect that the sentence is going be to an accurate description of the picture, despite the irrationality of such a procedure in the experimental context, where sentences are equally likely to be true and false. When this prediction is not borne out, this manifests as a negative deflection in the ERPs, presumably reflecting integration or retrieval difficulty, depending on ones' preferred interpretation of N400-family effects (Baggio & Hagoort, 2011; Brouwer et al., 2012; Brown & Hagoort, 1993; Hagoort et al., 2004; Kutas & Federmeier, 2011; Lau et al., 2009, 2008; Nieuwland et al., 2020). Whether the effect that we observe is an early onset N400 (Van Petten, Coulson, Rubin, Plante, & Parks, 1999; Vissers, Kolk, van de Meerendonk, & Chwilla, 2008) or an N2b (D'Arcy et al., 2000; Knoeferle et al., 2011; Wassenaar & Hagoort, 2007) reflecting perceptual mismatch between the representation of the picture and the sentence, is inconsequential to the aptitude of such a sentence processing model. Since both these interpretations require knowing the truth value of the sentence – in the N2b case, in order to realize that the sentence does not match the picture – for the present purposes, they amount to the same thing.

On this model of sentence processing, the verification happens incrementally, as soon as new information becomes available. Upon the completion of the subject noun phrase, participants start building a model that will make the sentence true of the preceding picture: if the claim is about 'most of the circles', the participants immediately try to figure out what predicate would make the sentence true, thus generating predictions about the final word. Because the verification procedure is instantiated before the participants know the truth value of the sentence, the verificational differences associated with the quantifier classes are instantiated at least as early as the noun. As mentioned in chapter 3, the host of differences in frequency and length etc., would make comparisons at the quantifier unreliable, but it is entirely possible that the differences manifest as early as this as well. This would also explain why truth value effects do not arise until the truth value is known (e.g Augurzky et al., 2017): because the true model of the sentence is not yet falsified, effects of violated predictions do not occur.

If one is not inclined to accept such a model of sentence processing, one might attempt a different explanation of the effect, for example one in which the negativity is in reality a decision effect, akin to what I outline below for the positivity following the N400-esque activity for the False versus True comparison for proportional quantifiers. So, in this model, you are not building a model and thus have no predictions for the final word. The negativity is then explained by the fact that deciding that a sentence is false is more strenuous to the participants. There is some evidence in the literature that false sentences (Carpenter & Just, 1975; Chang, 1986) – and even negative sentences in general (Deschamps et al., 2015; Just & Carpenter, 1971) – are associated with longer reaction times. This is true, even in the context of the experiments presented herein. However, then you would firstly have to assume that false sentences are somehow special. If you are not building a model, it is not immediately obvious why that should be the case. The model building approach provides an explanation – violation of prediction – but if participants are not building a model, one should – in the absence of an alternative account of the differences – expect that true and false sentences are symmetric, since the only difference between them is their truth and falsity relative to the model. The challenge to reconcile their account with the data is therefore on this alternative account. More problematically, even on this approach, the difference between true and false sentences relies on knowing the truth value of the sentence. It seems implausible given the sentence internal effects we observe, that the verification procedure is not initiated until the final word, and any alternative explanation must account for these differences in some other way. We have already established that these effects are not related to syntacto-semantic composition, since they do not appear in the comprehension experiment. Because the nouns were identical across conditions, we can also rule out frequency effects. A different sentence processing model must therefore come up with an aspect of processing, not related to verification, that could explain the sentence internal differences.

### 5.3.2 Interaction with Computational Complexity

Interestingly, the sentence final effects explained by the proposed sentence processing model above is modulated by quantifier class. In particular, the difference between false and true sentences is smaller for proportional quantifiers, both statistically and in terms of voltage values, and is followed by a positivity in the explicit verification experiments, both in paper 1 and paper 2.[3] Additionally, there are no sentence final effects for proportional quantifiers in the comprehension experiment in paper 1, despite the fact that non-proportional quantifiers exhibit similar, albeit weaker, effects as those observed in the explicit verification experiments.

In light of the processing model, this suggests that the increased difficulty, i.e. the recruitment of additional memory resources, for proportional quantifiers, disturb subsequent processing. A natural interpretation of this disturbance is that participants may have less resources to predict the final word, resulting in a smaller N400 (Delogu et al., 2017; Nieuwland et al., 2010; Nieuwland & Kuperberg, 2008; Nieuwland & van Berkum, 2006). This could be the result of neither completion of the sentence (‘red’ or ‘yellow’) being as dominant for participants at the time of final word onset, compared to the non-proportional case. The ensuing positivity might reflect increased decision complexity (Augurzky et al., 2017; Sassenhagen et al., 2014), an interpretation that is strengthened by the fact that this positivity is significant for 4 digits, but not for 2, in the memory load experiment in paper 2.

However, recent evidence (Aurnhammer et al., 2021) presents a unified interpretation. According to Aurnhammer et al. (2021, see also Brouwer, Crocker, Venhuizen, and Hoeks (2017)), both the N400 and the P600 is modulated by unexpected words. If the final word is more unexpected for proportional quantifiers – due to the higher complexity of the verification algorithm – this interpretation of the N400 and P600 components predict precisely this pattern. Since the P600 has been argued to reflect integration difficulty (Brouwer & Hoeks, 2013), and a spatiotemporal overlap between the components can cause reductions in one as a result of an increase in the other (Brouwer & Crocker, 2017), the less predicted final word might cause

[3] In paper 2, this positivity is only significant for 4 digits and overall, but not for 2 digits.

subsequent integration into the discourse context to be more difficult, thus yielding a larger P600 and an accompanying reduction in the N400.

The results presented herein cannot adjudicate between these two interpretations of the sentence final effects for proportional quantifiers: whether there is more uncertainty about the final word – i.e. neither alternative is strongly predicted – leading to a reduction in the N400 and an increase in decision complexity reflected in a subsequent positivity, or whether the lower expectancy of the final word causes an increase in the P600 at the expense of the N400.

It is also worth discussing how the complexity of the verification algorithm impacts sentence processing when verification is not task relevant. The waveform profiles for non-proportional quantifiers in the comprehension experiment in paper 1, are largely consistent with the same sentence processing model discussed above: false sentences elicit a negative deflection relative to true. However, this is not the case for proportional quantifiers, which do not present any significant effects (no sentence final effect of truth value and no sentence internal difference between proportional and non-proportional quantifiers). This suggests that participants are not verifying sentences with proportional quantifiers, even though they do for non-proportional.

Prompted by this discrepancy, the sentence processing model might be in need of a slight revision. If the truth-value is not readily available – i.e. if the verification algorithm requires resources that participants need to perform another task – participants are not able to ascertain whether the sentence is an accurate description of the picture or not. They are thus not able to predict the final word for proportional quantifiers, since their cognitive resources are devoted to the main task, i.e. paying attention to whether the sentence and/or the picture contains certain elements, and consequently have no expectations about what the final word is going to be. This suggests that the model building approach to sentence processing is affected by algorithmic complexity, such that participants are building a true model of the sentence only if the complexity of the verification algorithm is below a certain threshold, which might vary depending on the complexity of experimental task. In the event that the verification algorithm is too complex, they might instead build two separate models – one for the picture and one for the sentence – that they compare when required to do so, e.g. when asked a comprehension question about the picture and the sentence. Since knowing the truth value of the sentence was never necessary to respond to the comprehension question, the increased verification complexity never affected reaction times: the two models were only superficially compared, rather than being fully integrated.

## 5.4 Chapter summary

As a way of summarizing, I will repeat the research questions and answer them in turn:

(1) Do the differences in the computational complexity of verification algorithms for proportional and non-proportional quantifiers manifest in distinct brain responses?

(2) If so, are these brain responses related to memory, as predicted by the automata theory?

(3) At what point(s) during sentence processing do such differences occur?

(4) What model of sentence processing best explains such patterns?

(1) Brain responses to quantified sentences are affected by the complexity of the verification algorithm of the quantifier as formalized by the semantic automata theory, and (2) these differences are best explained by the deployment of memory resources, as evidenced by the recruitment of brain regions associated with working memory and executive functioning and ERP components associated with memory. Nevertheless, the nature of these memory resources depend on the task: in a simple picture-sentence verification task, a late positivity, sometimes associated with recollection memory, is found for proportional versus non-proportional quantifiers, whereas an early anterior negativity, traditionally associated with working memory during sentence processing, is found for the same comparison under memory load. What prompts the different concrete implementations of the abstract automaton memory in the human brain cannot be inferred from the extant experimental results, and further study is required to ascertain the factors influencing the differences observed between experiments.

(3) There is evidence of differential waveforms between proportional and non-proportional quantifiers at the earliest point in the sentence where the verification procedure could be isolated, i.e. the completion of the noun phrase. They might also be detectable earlier, but confounds of length, frequency, and morphosyntactic differences prevent valid conclusions to be drawn. However, these differences permeate throughout the entire sentence, influencing subsequent prediction and decision processes. (4) The results are best explained by a model of sentence processing where participants are actively predicting the unfolding sentence to match, i.e. be true of, the active representation of the picture. This is true regardless of whether participants are actively verifying the sentence, or whether verification is not necessary to perform the task. An important exception is that proportional quantifiers

do not trigger effects of truth value or verification when verification is not required to perform the task. The absence of these effects indicate that the complexity of the verification algorithm impacts the neural signals elicited by sentence processing, irrespective of the relevance of verification, and that limitations in processing resources compel participants to process sentences differently as a function of the interaction between task complexity and the complexity of verification.

# Chapter 6
# Outlook and Directions for Future Research

The previous chapter was devoted to drawing the conclusions it was possible to draw on the basis of previous studies and the experimental work presented in this thesis. This is of course not to say that this concludes the study of the processing of natural language quantifiers or the application of algorithmic analysis to sentence processing more generally. This final chapter is therefore devoted to the road ahead, the questions left open, and possible avenues to pursue for future research. Sections 6.1 and 6.2 deal with how the work generalizes and how its methodology can be applied to the study of other linguistic domains. I discuss what I believe to be the most fruitful avenues to pursue in future research in 6.3, before providing some concluding remarks in 6.4.

## 6.1 The Present Work as a Proof of Concept

The work presented in this thesis demonstrates that the computational complexity of the verification algorithm for natural language quantifiers is reflected in distinct neural responses. This compelling evidence suggests that human language processing is subjected to the same constraints as those applicable to abstract machines. Of particular interest is the fact that the nature of the difference we explored – requiring or not requiring some abstract notion of memory – manifests in the recruitment of concrete memory systems. An interesting open question, to be discussed in 6.3.1 below, is that the specific memory systems that are recruited seems to differ depending on the nature of the task. This, however, does not undermine the importance of the findings, since what matters is not the translation of the memory effect into concrete memory systems, but the simple fact that it is translated and the resultant corroboration of the formal generalizations.

The principle value of the present results lies in their application of predictions from theoretical computer science to electrophysiological data. Using complexity classes like those developed for quantifiers to approximate the algorithmic level, can help bridge the gap between the computational and the implementational level, thereby facilitating more integrative neurolinguistic and/or cognitive neuroscientific theories (Baggio et al., 2016, 2015; Embick & Poeppel, 2015; Isaac et al., 2014; S. Lewis & Phillips, 2015). Additionally, the mathematical precision of such complexity classes, has caused scholars to maintain the necessity of these analyses for all psychological theories (van Rooij & Baggio, 2020, 2021; van Rooij et al., 2019). The fact that these predictions are borne out in a case study on quantifier verification, indicates, firstly, that complexity theoretic analyses can explain brain responses, but also that brain responses that are not in line with the formal predictions warrant revision of the formal analysis, e.g., whether heuristics or approximation algorithms are involved (Carruthers et al., 2018; van Rooij & Wareham, 2012; van Rooij et al., 2012). However, as this thesis is primarily concerned with the application of the formal results and not with their derivation, I will focus on the former in what follows.

As mentioned in chapter 3, behavioral responses to problems in social cognition and deductive reasoning are explained as a function of their computational complexity (Gierasimczuk et al., 2013; Szymanik, Meijering, & Verbrugge, 2013; Zhao et al., 2018). An interesting avenue to pursue, given the results presented herein, is whether the same patterns hold true for brain responses. Brain responses could be a sensitive measure of the aptitude of hypothesized approximation algorithms, such as those proposed for the travelling salesperson problem (Carruthers et al., 2018; Graham et al., 2000). Also for language, the present results have interesting implications.

## 6.2 Implications for Other Linguistic Domains

Circularity is a recurring, and sometimes unavoidable, problem in semantics. The fact that a word or sentence meaning has to be defined in terms of other word or sentence meanings, creates a vicious cycle. This has sparked cross-linguistic research into defining semantic primitives from which all other concepts can be derived (e.g. Wierzbicka, 1996). Even setting aside the philosophical worry that all primitives are arbitrary (Goodman, 1977)[1], such endeavours have been subject to criticism. It may not be easily determinable which primitives are necessary and/or sufficient – i.e. where to stop decomposition – nor whether the meaning of every linguistic term corresponds to a unique intersection of such primitives (Jackendoff, 2002).[2]

[1] Since all terms can be defined in terms of each other, primitives are not privileged. While primitives are indefinable within a system – in virtue of being primitives – they can easily be defined outside of the system, using different primitives.

[2] As an historical example, it is worth mentioning generative semantics (for an overview see e.g. McCawley, 1995), and its critics (e.g. Fodor, 1970).

Such problems have caused certain researchers to abandon the search for primitives, and instead focus on the compositional nature of meaning even at the lexical level (Pustejovsky, 1995), or how the semantic, *qua* lingustic, system interacts with our knowledge of the world (Jackendoff, 2002).

Another option is to try to describe meanings in non-linguistic vocabulary, such as logic or mathematics. Since quantifiers are one of the few linguistic theories for which there is a successful non-linguistic theory, they have been utilized in analyzing a whole range of other linguistic constructions (Peters & Westerståhl, 2006). Aside from the most accepted extension into adverbial quantification – which has been shown to give rise to similar electrophysiological effects as their nominal counterparts (Augurzky, Hohaus, & Ulrich, 2020) – they have been used to analyze tense (Fernando, 2004, 2007) and aspect (ter Meulen, 1991), as well as attitude verbs (Moltmann, 2003) and modals (van der Hoek & de Rijke, 1993).

This means that there is a host of phenomena that, hypothetically, should behave similarly to quantifiers. Providing semantic automata for these quantifiers – as in, e.g., ter Meulen (1991) – would generate predictions about the complexity of their verification. Provided that their domain of quantification – e.g., times, events, worlds – could be visualized in a similar manner to the concrete objects denoted by ordinary determiners, these predictions could be tested in a comparable experimental paradigm. The pictoral nature of Fernando's (2004; 2007) languages for tense is a viable candidate for such designs.

One potential problem with many of these other uses is that their resulting formal languages are all regular, and the differences we have observed are between FSAs and PDAs, i.e. between regular and context-free languages. While other measures of complexity have been used, such as approximate Kolmogorov complexity (van de Pol et al., 2019) or minimal description length (van de Pol, Lodder, van Maanen, Steinert-Threlkeld, & Szymanik, 2021), their translation into concrete cognitive predictions is less direct. Similar concerns apply to the subregular distinctions discussed in 6.3.3 below, that the reader may recall from 2.2.1 concern further decomposition of complexity within the regular region of the Chomsky hierarchy. However, these are directions worth exploring both from a formal and an experimental angle.

In fact, recent work in *artificial grammar learning* (AGL) using ERPs, has investigated the learning of phonological patterns, that have been shown to be subregular (Heinz, 2018; Heinz & Rawal, 2011), to explore precisely such issues (Avcu, 2019; Avcu, Rhodes, & Hestvik, 2019; Tsogli, Jentschke, & Koelsch, 2022). While complexity have not been explicitly manipulated within experiments, varying degrees of complexity have impacted results: when *strictly local* complexity increases, but *strictly piecewise* complexity does not, participants did not show behavioral or neural signs of learning (Avcu, 2019).[3] This highlights some methodological problems – i.e. are participants solving the correct learning problem – but also demonstrates that, at the very least, there are indirect measures of the complexity of computational problems that are detectable in the evoked potential. As a consequence, if the formal specification of learning problems (e.g. Heinz, 2016; Niyogi, 2006) associates different complexity profiles with learning two distinct phonological patterns, it is in principle possible to generalize the theoretical foundations of the present project to radically different domains of linguistics, such as phonology.

## 6.3 Outstanding Questions

There are, however, still questions pertaining to quantifier verification that remains unanswered. Most immediate of these is the relationship between the abstract notion of memory from the automata theory and its precise implementation in human brains. This will be discussed in 6.3.1. The remaining two sections will be an attempt to connect with the two research programs that this project falls somewhat in between. On one side, 6.3.2 deals mostly with the highly experimental ventures of procedural semantics and the search for canonical verification procedures for certain quantifier expressions, and, more generally, with the integration of such algorithms into the wider context of cognition. On the other, 6.3.3 deals with the decidedly theoretical refinement of the analysis of quantifier languages into distinct subclasses of regular languages, and consequently of automata types.

### 6.3.1 The Nature of Memory

Echoing Niyogi (2006), formal algorithmic results are unlikely to accurately represent cognitive reality, but the insight gained from their mathematical certainty is profound. It is therefore not surprising that different memory systems are involved during quantifier verification, depending on the nature and/or complexity of the task. However, being content with the mathematical insight is arguably an instance of intellectual indolence, and further experimentation and computational modelling could help shed light on what causes the differences in the evoked potential between the two experiments.

Recall from chapter 5 that one remaining open problem is whether the concrete implementation of the abstract automata stack memory overlaps with recollection memory or whether the two are distinct. I suggested in 5.2, that the results from paper 1 are consistent both with an interpretation where the component modulated by proportional quantifiers is a *late positive component* (LPC), related to recollection (Rugg & Curran, 2007), or a *positive slow wave*, argued to index retrieval from

[3] See e.g. De Santo and Rawski (2022) or Rogers et al. (2013) for an introduction to different subregular languages.

short-term memory (García-Larrea & Cézanne-Bert, 1998). The fact that we find an early *left anterior negativity* (LAN) or *sustained anterior negativity* (SAN) when modulating memory load in paper 2, seems to favor an interpretation where the effect in paper 1 is an LPC, and that the effect in paper 2 is related to working memory. However, this interpretation is in conflict with previous results that do not involve recollection, but elicit comparable responses (De Santo et al., 2019; McMillan et al., 2005, 2006; Morgan et al., 2011; Olm et al., 2014).

It is therefore pertinent to manipulate the complexity of recollection. This could be done in a paradigm similar to that applied in paper 1, with the addition of recollection complexity as an independent variable. This could involve increasing the number of objects in the picture preceding the trial, either in the form of more tokens or more types; removing the 2x2 grid structure for half of the trials; or having half the experiment conducted in a similar fashion to those presented in the thesis, with the same picture before every trial in a block, and having a different picture before every trial in the other half. If the differences are not affected by manipulating recollection complexity, then the component manipulated is plausibly a PSW. If, on the other hand, there is an interaction between quantifier class and recollection complexity, then a case can be made that the component is an LPC.

However, this is intertwined with another open problem. The differences we observe between experiments – i.e. a late positivity with no additional memory task and an early negativity under memory load – seem to be driven by task effects. As discussed in 5.1, this can be explained in two ways: Either (1) the early negativity obscures a later positivity, or (2) the complexity of the task forces other working memory systems to be recruited. (1) seems to presuppose that the positivity is related to recollection, but could also be the result of a modular working memory (Baddeley, 2012). More problematically, its reliance on unobservables – i.e. the concealed positivity – makes it very hard to test. A more fruitful approach would therefore be to pursue (2). If one is able to ascertain in which situations the two components are evoked, one might be able to gauge what memory systems are engaged and what triggers the switch from one to the other. We already know that the presence of a digit matching task alters the evoked potential, so a likely first avenue is to manipulate other working memory subsystems, such as the visuospatial scratchpad, by, e.g., making participants remember the location of a dot.

Summing up, it is timely to point out that their platonic reality notwithstanding, formal results will only get you so far. At the level of neurobiological systems, additional assumptions are required to make specific predictions. These assumptions could be put to the test by systematic alternations of the paradigm presented in this thesis. However, there are also other possiblities to consider.

### 6.3.2 Procedural Semantics

As mentioned in the introductory chapter, *procedural semantics* views the meaning of declarative sentences as the set of algorithms computing the extension of the sentence, i.e. its truth value (Moschovakis, 2006; Muskens, 2005; Pietroski et al., 2011; Suppes, 1982; Szymanik, 2016; Tichý, 1969; van Benthem, 1986b; van Lambalgen & Hamm, 2005). The experiments presented herein were not designed to test this philosophical position, but it is pertinent to point out that verification effects, at least for proportional quantifiers, seem to be reserved for explicit verification tasks. That being said, the model of sentence processing I argued for in chapter 5 is very much in line with the procedural semantics view. I would therefore like to suggest how the work presented herein can inform, as well as be informed by, procedural semantics.

The representations upon which the algorithms are hypothesized to operate, greatly affect the issues of complexity discussed in the present work. As mentioned in chapters 1 and 2, the semantic automata rely on examining each and every object denoted by the quantified noun phrase sequentially. However, the psychological literature on quantity perception has known for a long time that this is not how humans perceive quantities. Small numbers are recognized immediately through a process known as *subitizing* (e.g. Brysbaert, 2018; Feigenson, Dehaene, & Spelke, 2004), whereas large numbers are approximated using the *Approximate Number System* (ANS) (see e.g. Dehaene, 2011; Odic & Starr, 2018). Importantly, none of these processes involve counting or inspecting objects sequentially. It has been shown that behavioral measures are affected by the mode of presentation (Knowlton et al., 2021; Lidz et al., 2011; Pietroski et al., 2009; Steinert-Threlkeld, Munneke, & Szymanik, 2015), such that, e.g., the verification of ‘more than half’, but not that of ‘most’, benefits from paired visual stimuli, i.e. that all B As and all non-B As are immediately next to each other, with the remainder (e.g. the As that are B if ‘more than half’ is true) being by themselves. This can be explained by the algorithm on paired stimuli being computable by a finite-state automaton (FSA), whereas the non-paired requires a pushdown automaton (PDA) (Steinert-Threlkeld et al., 2015). This means that one’s conception of a visual stimuli can influence the complexity of the algorithm.

Conversely, it has been argued that the semantics of certain quantifiers can impact one’s conception of a visual scence (Hackl, 2009; Knowlton, Pietroski, Halberda, & Lidz, 2022; Knowlton, Trueswell, & Papafragou, 2022; Lidz et al., 2011). Fuelled by the procedural semantics idea to describe the set of algorithms computing the extension of quantified expressions, researchers have attempted to associate certain kinds of algorithms with specific quantifiers. In particular, Lidz et al. (2011) coined the *Interface Transparency Thesis* (ITT), according to which speakers are biased towards a verification algorithm that mirrors the canonical specification of an expression's truth conditions. Of particular interest are expressions that are logically equivalent, but are seemingly associated with different verification strategies. The original example is 'most' and 'more than half' (Hackl, 2009; Talmina et al., 2017).[4] The fact that 'most' is not affected by paired stimuli (Lidz et al., 2011; Steinert-Threlkeld et al., 2015) has been explained by 'most', rather than being a precise notion of more than half, relying on the ANS to make an estimate as to which of the subsets $A \cup B$ and $A - B$ is the largest. More recent work has also sought to attribute differences in the verification of universal quantifiers to algorithms operating on first- or second-order predicates (Knowlton, Pietroski, et al., 2022; Knowlton, Trueswell, & Papafragou, 2022), or, for proportional quantifiers 'most' and 'more', differences in their set-theoretic description (Knowlton et al., 2021).

Starting with how the present work can be informed by procedural semantics, the two experimental avenues above can perhaps shed light on some seeming inconsistencies in the experimental data. Rather than being related to the recruitment of distinct memory systems, it might be worth noting that the proportional quantifiers in paper 2 were different to those in paper 1. Paper 2 explicitly targeted 'more/less than half' and 'majority/minority', as well as the comparative version of 'most' and 'fewest' ('flest/færrest av'), rather than the proportional 'de fleste' and compartive 'færrest av' from paper 1. Seeing as these have been found to give rise to different verification strategies (Hackl, 2009; Knowlton et al., 2021; Lidz et al., 2011; Pietroski et al., 2009; Steinert-Threlkeld et al., 2015; Talmina et al., 2017), perhaps the different ERP components that were recruited could be explained by this. In fact, Steinert-Threlkeld et al. (2015) found that only 'more than half' interacted with working memory load in their experiment. Assuming that 'most' relies heavily on the ANS and that only 'more than half' involves counting, it is possible that the effect in paper 1 is recollection of the visual stimulus in order to estimate the ratio between red and yellow circles or triangles, whereas the effect in paper 2 is an effect of counting the objects stored in short-term memory. In that event, it is not surprising that we observe an LPC in paper 1 and a SAN or LAN in paper 2. An experiment could easily be designed to test this hypothesis: it only requires that there be enough trials with 'most' and 'more than half' to be able to compare them directly.

The nature of the pictorial stimuli could also be put under scrutiny. For simplicity's sake, the images in the present experiment consisted of grouped shapes that, within each quadrant, were subitizable. Since this was a first attempt at detecting

[4] But see Denić and Szymanik (2022) for an argument that they are not truth-conditionally equivalent.

complexity differences, the most important consideration was to ensure that participants could verify the images. Considering that different quantifiers seemingly interact with magnitude representations differently, subsequent studies should aim to test whether randomized or paired stimuli impact the evoked potenital. Of particular interest is the fact that the ‘more than half’ algorithm operating on pairs is computable by an FSA (Steinert-Threlkeld et al., 2015), and should therefore not recruit the memory resources associated with PDAs. This would provide another avenue that can explore the neuropsychological consequences of the choice of verification algorithm, and could thus inform the research program of procedural semantics. One interesting open problem is the interaction between quantifier verification and the ANS (Knowlton et al., 2021; Knowlton, Pietroski, et al., 2022; Lidz et al., 2011), which should be informed by the complexity theoretic notions of approximation algorithms (Carruthers et al., 2018; van Rooij & Wareham, 2012; van Rooij et al., 2012).

### 6.3.3 Subregularity and Permutation Closure

Going into further detail, it is also worth mentioning that a growing body of literature has been examining more fine-grained notions of the complexity of string languages and their associated automata. As mentioned in 2.2.1, many scholars have come to hypothesize *subregularity* – i.e. being a proper subset of regular languages – to be a unifying feature of human language (Chandlee, 2017; De Santo & Rawski, 2022; Graf, 2012, 2017, 2019; Heinz, 2018; Heinz & Rawal, 2011). Despite previously having had its primary application in phonology and syntax, subregular analyses have recently been applied to generalized quantifiers. Motivated by the counterintuitive fact that, e.g., ‘an even number of’ and ‘all’ have the same computational complexity, i.e. regular, Graf (2019) demonstrates that Aristotelian and numerical quantifiers are all *tier-based strictly local* (TSL), whereas parity quantifiers are not. He further subdivides the TSL quantifiers into *monotonic* and *non-monotonic TSL*, where only the former type seems to be instantiated as lexical determiners in natural languages.

This demonstrates that there is more nuance in the complexity of generalized quantifiers that could potentially be explored, and leaves the open question of whether such nuances are detectable at the neural level. As mentioned in 6.2 above, the learnability of subregular patterns has been demonstrated to give rise to electrophysiological effects in AGL experiments, but these do have some methodological problems (Avcu, 2019): what type of patterns are subjects looking for when they are learning a pattern? Some patterns are very complicated if they are viewed as *strictly local* (SL), but might turn out to be very simple *strictly piecewise* (SP) or TSL. One advantage of quantifier verification over AGL, is that, as demonstrated in this thesis, the minimal complexity of the verification algorithm has been shown to manifest in the evoked potential. A related phenomenon is the hypothesis that numerical quantifiers should increase with higher numbers (Szymanik, 2016), which does have some support in behavioral data (Szymanik & Zajenkowski, 2010b), or the fact that parity quantifiers require cyclic FSAs, whereas Aristotelian and numerical quantifiers do not. Testing whether such minor differences between subsets of FSAs (see McNaughton & Papert, 1971; Rogers et al., 2013, for examples of such subsets) are reflected in detectable neural responses is one potential follow-up to this study. However, as discussed in 6.2 above, the differences between such abstract machines have less clear cognitive counterparts; the difference between requiring or not requiring memory is a clear cognitive difference in a way a cyclic or acyclic automaton is not. On the other hand, it could shed light on the nature of the differences observed in present experiment, by adjudicating between interpretations of the ERP components: if similar difference waves are observed between cyclic and acyclic FSAs and the difference between FSAs and PDAs in one of the experiments, this is an argument that it should be viewed as a generic processing cost rather than specifically memory.

Lastly, a final point made by Graf (2019) ties in with the discussion in 6.3.2. Recall that generalized quantifiers are, by definition, permutation invariant. However, if quantifier languages are not required to be permutation invariant, it turns out that the languages corresponding to proportional quantifiers such as 'most' or 'half', can be described as a monotonic TSL language, and they are thus computable by FSAs. Without going into the formalities, if you are not allowed to start or end with non-B As for a quantifier $Q(A, B)$, 'most' is true if there are no consecutive non-B As, i.e. for every non-B A there is at least one B A. This mirrors the pairing effect for 'more than half' (Lidz et al., 2011; Steinert-Threlkeld et al., 2015), and suggests that there are interesting parallels between purely theoretical definitions of patterns in the subregular complexity literature and the attempt in psycholinguistics to place representations of semantic information within the wider systems of cognition. If anything is to be learned from the present thesis, it is precisely the reciprocal benefit of integrating formal computer science and psycholinguistics.

## 6.4 Concluding Remarks

I envision this thesis to be the first step on the path of a new research program that explores the neural consequences of formal algorithmic results. While there are answers and borne out predictions to be found, the doors that are opened by these answers are, potentially, far more interesting. For quantifier verification specifically, it will be worthwhile to explore task effects as well as the impact of the choice of quantifiers and/or visual stimuli. This will allow us to describe how the abstract notions of computational complexity, such as stack-memory, manifest in the recruitment of concrete memory resources in human subjects, as well as how algorithms of quantifier verification interact with other cognitive systems such as the magnitude processing system. It might additionally provide an avenue for exploring whether more fine-grained distinctions, such as subregular languages and their corresponding subsets of finite automata, also manifest as detectable neural effects. If they are, this opens up a host of other research areas such as examining the learning of phonological patterns, or testing whether other semantic phenomena analyzed in terms of generalized quantifiers or formal languages, are apt descriptions of their cognitive reality.

An alternative avenue to explore, is whether different kinds of complexity theoretic analysis have the same predictive power as the minimal complexity of a computational problem. This concerns, e.g., whether algorithms are computable in linear, logarithmic or exponential time, and applies to problems such as the travelling salesperson problem.

However, for both the more and the less fine-grained analyses of computational problems and their associated algorithms, the challenge to the researcher is to conceptualize what their parallel in human cognition should be. As an example, consider algorithms that are computable in linear time. These predict a linear increase in reaction time, but the consequences for cognitive resources are much more opaque. Since every subsystem must have a specified role for mechanistic explanation to be possible, more theoretical work is needed in order for the translation between computer science and cognitive science to be fruitful.

# Bibliography

Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2019). Measuring the cognitive cost of downward monotonicity by controlling for negative polarity. *Glossa: a journal of general linguistics*, *4*. doi: 10.5334/gjgl.770

Ahn, D., & Sauerland, U. (2017). Measure constructions with relative measures: Towards a syntax of non-conservative construals. *The Linguistic Review*, *34*, 215-248. doi: 10.1515/tlr-2017-0001

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2014). Inhibition and the right inferior frontal cortex: One decade on. *Trends in Cognitive Sciences*, *18*(4), 177–185. doi: 10.1016/j.tics.2013.12.003

Augurzky, P., Bott, O., Sternefeld, W., & Ulrich, R. (2017). Are all the triangles blue? - ERP evidence for the incremental processing of German quantifier restriction. *Language and Cogntion*, *9*, 603–636. doi: 10.1017/langcog.2016.30

Augurzky, P., Hohaus, V., & Ulrich, R. (2020, 11). Context and Complexity in Incremental Sentence Interpretation: An ERP Study on Temporal Quantification. *Cognitive Science*, *44*(11). doi: 10.1111/cogs.12913

Augurzky, P., Schlotterbeck, F., & Ulrich, R. (2020, 11). Most (but not all) quantifiers are interpreted immediately in visual context. *Language, Cognition and Neuroscience*, *35*(9), 1203–1222. doi: 10.1080/23273798.2020.1722846

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLOS ONE*, *16*. doi: 10.1371/journal.pone.0257430

Avcu, E. (2019). *Using cognitive neuroscience to understand learning mechanisms: Evidence from phonological processing* (Unpublished doctoral dissertation). University of Delaware.

Avcu, E., Rhodes, R., & Hestvik, A. (2019). Neural Underpinnings of Phonotactic Rule Learning. In K. Hout, A. Mai, A. McCollum, S. Rose, & M. Zaslansky (Eds.), *Proceedings of the 2018 annual meeting on phonology.* Washington, DC: Linguistic Society of America. doi: 10.3765/amp.v7i0.4487

Axel, M., & Müller, N. G. (1996). Dissociations in the Processing of "What" and "Where" Information in Working Memory: An Event-Related Potential Analysis. *Journal of Cognitive Neuroscience*, *8*, 453–473. doi: 10.1162/jocn.1996.8.5.453

Bach, E., Jelinek, E., Kratzer, A., & Partee, B. H. (Eds.). (1995). *Quantification in Natural Languages.* Dordrecht: Kluwer Academic Publishers. doi: 10.1007/978-94-017-2817-1

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, *63*, 1–29. doi: 10.1146/annurev-psych-120710-100422

Badre, D., & Wagner, A. D. (2004). Selection, Integration, and Conflict Monitoring: Assessing the Nature and Generality of Prefrontal Cognitive Control Mechanisms. *Neuron*, *41*, 473–487. doi: 10.1016/S0896-6273(03)00851-1

Baggio, G. (2018). *Meaning in the brain.* Cambridge, MA: MIT Press.

Baggio, G. (2020). Epistemic Transfer Between Linguistics and Neuroscience: Problems and Prospects. In R. Nefdt, C. Klippi, & B. Karstens (Eds.), *The philosophy and science of language: Interdisciplinary perspectives* (pp. 275–308). Cham: Palgrave Macmillan. doi: 10.1007/978-3-030-55438-5_11

Baggio, G. (2021). Compositionality in a parallel architecture for language processing. *Cognitive Science*, *45*(5), e12949. doi: 10.1111/cogs.12949

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N40. *Language and Cognitive Processes*, *26*, 1338–1367. doi: 10.1080/01690965.2010.542671

Baggio, G., Stenning, K., & van Lambalgen, M. (2016). Semantics and cognition. In M. Aloni & P. Dekker (Eds.), *The cambridge handbook of formal semantics* (pp. 756–774). Cambridge: Cambridge University Press.

Baggio, G., van Lambalgen, M., & Hagoort, P. (2008). Computing and recomputing discourse models: An ERP study. *Journal of Memory and Language*, *59*, 36–53. doi: 10.1016/j.jml.2008.02.005

Baggio, G., van Lambalgen, M., & Hagoort, P. (2015). Logic as marr's computational level: Four case studies. *Topics in Cognitive Science*, *7*(2), 287–298.

Barwise, J. (1979). On branching quantifiers in English. *Journal of Philosophical Logic*, *8*, 47–80. doi: 10.1007/BF00258419

Barwise, J., & Cooper, R. (1981). Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, *4*, 159–219.

Bastiaansen, M., Mazaheri, A., & Jensen, O. (2012). Beyond ERPs: Oscillatory Neuronal Dynamics. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford Handbook of Event-Related Potentials.* Oxford: Oxford University Press.

Bechtel, W. (2005). The Challenge of Characterizing Operations in the Mechanisms

Underlying Behavior. *Journal of the Experimental Analysis of Behavior*, *84*, 313–325. doi: 10.1901/jeab.2005.103-04

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A Mechanistic Alternative. *Studies in History and Philosophy of the Biological and Biomedical Sciences*, *36*, 421–441. doi: 10.1016/j.shpsc.2005.03.010

Bird, A. (2021). Understanding the Replication Crisis as a Base Rate Fallacy. *The British Journal for the Philosophy of Science*, *74*, 965–993. doi: 10.1093/bjps/axy051

Bogen, J. (2005). Regularities and causality; generalizations and causal explanations. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*, 397–420. doi: 10.1016/j.shpsc.2005.03.009

Boone, W., & Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese*, *193*, 1509–1534. doi: 10.1007/s11229-015-0783-4

Bremnes, H. S., Szymanik, J., & Baggio, G. (2022). Computational complexity explains neural differences in quantifier verification. *Cognition*, *223*, 105013. doi: 10.1016/j.cognition.2022.105013

Brouwer, H., & Crocker, M. W. (2017). On the Proper Treatment of the N400 and P600 in Language Comprehension. *Frontiers in Psychology*, *8*. doi: 10.3389/fpsyg.2017.01327

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*, *41*, 1318–1352. doi: 10.1111/cogs.12461

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143. doi: 10.1016/j.brainres.2012.01.055

Brouwer, H., & Hoeks, J. C. (2013). A time and place for language comprehension: mapping the N400 and the P600 to a minimal cortical network. *Frontiers in Human Neuroscience*, *7*, 758. doi: 10.3389/fnhum.2013.00758

Brown, C., & Hagoort, P. (1993). The Processing Nature of the N400: Evidence from Masked Priming. *Journal of Cognitive Neuroscience*, *5*, 34–44. doi: 10.1162/jocn.1993.5.1.34

Brunoni, A. R., & Vanderhasselt, M.-A. (2014). Working memory improvement with non-invasive brain stimulation of the dorsolateral prefrontal cortex: A systematic review and meta-analysis. *Brain and Cognition*, *86*, 1–9. doi: 10.1016/j.bandc.2014.01.008

Brysbaert, M. (2018). Numbers and Language: What's New in the Past 25Years? In A. Henik & W. Fias (Eds.), *Heterogeneity of function in numerical cognition* (pp. 3–26). London: Academic Press. doi: 10.1016/B978-0-12-811529-9.00001-7

Buxton, R. B. (2009). *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques* (2nd ed.). Cambridge: Cambridge University Press.

Buxton, R. B., Uludağ, K., Dubowitz, D. J., & Liu, T. T. (2004). Modeling the hemodynamic response to brain activation. *NeuroImage*, *23*, S220–S233. doi: 10.1016/j.neuroimage.2004.07.013

Carcassi, F., Steinert-Threlkeld, S., & Szymanik, J. (2021). Monotone Quantifiers Emerge via Iterated Learning. *Cognitive Science*, *45*. doi: 10.1111/cogs.13027

Carcassi, F., & Szymanik, J. (2021). An alternatives account of 'most' and 'more than half'. *Glossa: a journal of general linguistics*, *6*. doi: 10.16995/glossa.5764

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, *82*, 45–73. doi: 10.1037/h0076248

Carruthers, S., Stege, U., & Masson, M. E. J. (2018). The Role of the Goal in Solving Hard Computational Problems: Do People Really Optimize? *Journal of Problem Solving*, *11*. doi: 10.7771/1932-6246.1200

Chandlee, J. (2017). Computational locality in morphological maps. *Morphology*, *27*, 599–641. doi: 10.1007/s11525-017-9316-9

Chang, T. M. (1986). Semantic Memory: Facts and Models. *Psychological Bulletin*, *99*, 199–220. doi: 10.1037/0033-2909.99.2.199

Chemla, E., Dautriche, I., Buccola, B., & Fagot, J. (2019). Constraints on the lexicons of human languages have cognitive roots present in baboons (Papio papio). *PNAS*, *116*(30). doi: 10.17605/OSF.IO/U72H3.y

Chemla, E., & Singh, R. (2014). Remarks on the Experimental Turn in the Study of Scalar Implicature, Part II. *Language and Linguistics Compass*, *8*, 387–399. doi: 10.1111/lnc3.12080

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, *2*, 113–124. doi: 10.1109/TIT.1956.1056813

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: The MIT Press.

Clark, H. H. (1976). *Semantics and Comprehension*. The Hague: Mouton.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472–517. doi: 10.1016/0010-0285(72)90019-9

Clark, H. H., & Chase, W. G. (1974). Perceptual coding strategies in the formation and verification of descriptions. *Memory and Cognition*, *2*, 101–111. doi:

10.3758/BF03197499

Coppock, E. (2019). Quantity Superlatives in Germanic, or "Life on the Fault Line Between Adjective and Determiner". *Journal of Germanic Linguistics*, *31*, 109–200. doi: 10.1017/S1470542718000089

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Cresswell, M. (1976). The semantics of degree. In B. H. Partee (Ed.), *Montague grammar* (pp. 261–292). New York: Academic Press.

Cummins, C., & Katsos, N. (Eds.). (2019). *The Oxford Handbook of Experimental Semantics and Pragmatics*. Oxford: Oxford University Press.

D'Arcy, R. C. N., Connolly, J. F., & Crocker, S. F. (2000). Latency shifts in the N2b component track phonological deviations inspoken words. *Clinical Neurophysiology*, *111*, 40-44. doi: 10.1016/S1388-2457(99)00210-2

de Groot, A. M. B., & Hagoort, P. (Eds.). (2018). *Research methods in psycholinguistics and the neurobiology of language: A practical guide*. Oxford: John Wiley & Sons.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*, 1–42. doi: https://doi.org/10.1016/0010-0277(92)90049-N

Dehaene, S. (2011). *The Number Sense: How the Mind Creates Mathematics* (2nd ed.). Oxford: Oxford University Press.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003, 5). Three parietal circuits for number processing. *Cognitive Neuropsychology*, *20*(3-6), 487–506. doi: 10.1080/02643290244000239

Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, *135*. doi: 10.1016/j.bandc.2019.05.007

Delogu, F., Crocker, M. W., & Drenhaus, H. (2017). Teasing apart coercion and surprisal: Evidence from eye-movements and ERPs. *Cognition*, *161*, 46–59. doi: 10.1016/j.cognition.2016.12.017

de Mey, S. (1991). 'Only' as a Determiner and as a Generalized Quantifier. *Journal of Semantics*, *8*, 91–106. doi: 10.1093/jos/8.1-2.91

Denić, M., & Szymanik, J. (2022). Are Most and More Than Half Truth-Conditionally Equivalent? *Journal of Semantics*, *39*, 261–294. doi: 10.1093/jos/ffab024

De Santo, A., & Drury, J. E. (2020). Encoding and Verification Effects of Generalized Quantifiers on Working Memory. In Ö. Eren, A. Giannoula, S. Gray, C.-D. Lam, & A. M. Del Rio (Eds.), *Proceedings of the fifty-fifth annual meeting of the chicago linguistic society* (pp. 103–114). Chicago, IL: Chicago Linguistic Society.

De Santo, A., & Rawski, J. (2022). Mathematical Linguistics and Cognitive Complexity. In M. Danesi (Ed.), *Handbook of Cognitive Mathematics.* Cham: Springer. doi: 10.1007/978-3-030-44982-7_16-2

De Santo, A., Rawski, J., Yazdani, A. M., & Drury, J. E. (2019). Quantified Sentences as a Window into Prediction and Priming: An ERP Study. In E. Ronai, L. Stigliano, & Y. Sun (Eds.), *Proceedings of the fifty-fourth annual meeting of the chicago linguistic society* (pp. 85–98). Chicago, IL: Chicago Linguistic Society.

Deschamps, I., Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition*, *143*, 244–253. doi: 10.1016/j.cognition.2015.06.006

Dry, M., Lee, M. D., Vickers, D., & Hughes, P. (2006). Human Performance on Visually Presented Traveling Salesperson Problems with Varying Numbers of Nodes. *Journal of Problem Solving*, *1*, 20–32. doi: 10.7771/1932-6246.1004

Dry, M., Preiss, K., & Wagemans, J. (2012). Clustering, Randomness, and Regularity: Spatial Distributions and Human Performance on the Traveling Salesperson Problem and Minimum Spanning Tree Problem. *Journal of Problem Solving*, *4*. doi: 10.7771/1932-6246.1117

Embick, D., & Poeppel, D. (2015, 4). Towards a computational(ist) neurobiology of language: correlational, integrated and explanatory neurolinguistics. *Language, Cognition and Neuroscience*, *30*(4), 357–366. doi: 10.1080/23273798.2014.980750

Farkas, D., & Kiss, K. (2000). On the comparative and absolute readings of superlatives. *Natural Language and Linguistic Theory*, *18*, 417–455. doi: 10.1023/A:1006431429816

Fauconnier, G. (1975). Pragmatic Scales and Logical Structure. *Linguistic Inquiry*, *6*, 353–375. Retrieved from `https://www.jstor.org/stable/4177882`

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*, 307–314. doi: 10.1016/j.tics.2004.05.002

Fernando, T. (2004). A Finite-state Approach to Events in Natural Language Semantics. *Journal of Logic and Computation*, *14*, 79–92. doi: 10.1093/logcom/14.1.79

Fernando, T. (2007). Finite-State Descriptions For Temporal Semantics. In H. Bunt & R. Muskens (Eds.), *Computing meaning* (Vol. 3, pp. 347–368). Dordrecht: Springer. doi: 10.1007/978-1-4020-5958-2_14

Fiebach, C. J., Schlesewsky, M., & Friederici, A. D. (2001). Syntactic Working Memory and the Establishment of Filler-Gap Dependencies: Insights from ERPs and fMRI. *Journal of Psycholinguistic Research*, *30*, 321–338. doi: 10.1023/A:1010447102554

Fitch, W. T. (2014). Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition. *Physics of Life Reviews*, *11*, 329–364. doi: 10.1016/j.plrev.2014.04.005

Fodor, J. A. (1970). Three Reasons for Not Deriving "Kill" from "Cause to Die". *Linguistic Inquiry*, *1*, 429–438. Retrieved from `https://www.jstor.org/stable/4177587`

Freunberger, D., & Nieuwland, M. S. (2016, 9). Incremental comprehension of spoken quantifier sentences: Evidence from brain potentials. *Brain Research*, *1646*, 475–481. doi: 10.1016/j.brainres.2016.06.035

Fritz, I., & Baggio, G. (2020). Meaning composition in minimal phrasal contexts: distinct erp effects of intensionality and denotation. *Language, Cognition and Neuroscience*, *35*(10), 1295–1313. doi: 10.1080/23273798.2020.1749678

Fritz, I., & Baggio, G. (2021). Neural and behavioural effects of typicality, denotation and composition in an adjective–noun combination task. *Language, Cognition and Neuroscience*, 1–23. doi: 10.1080/23273798.2021.2004176

García-Larrea, L., & Cézanne-Bert, G. (1998). P3, Positive slow wave and working memory load: a study on the functional correlates of slow wave activity. *Electroencephalography and Clinical Neurophysiology*, *108*, 260–273. doi: 10.1016/S0168-5597(97)00085-3

Gärdenfors, P. (Ed.). (1987). *Generalized Quantifiers: Linguistic and Logical Approaches.* Dordrecht: D. Reidel Publishing Company. doi: 10.1007/978-94-009-3381-1

Garfinkel, A. (1981). *Forms of Explanation.* New Haven, CT: Yale University Press.

Gazdar, G., & Pullum, G. K. (1985). Computationally Relevant Properties of Natural Languages and Their Grammars. *New Generation Computing*, *3*, 273–306.

Gehring, W. J., Liu, Y., Orr, J. R., & Carp, J. (2012). The Error-Related Negativity (ERN/Ne). In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford Handbook of Event-Related Potentials.* Oxford: Oxford University Press.

Geurts, B., & van der Silk, F. (2005). Monotonicity and Processing Load. *Journal of Semantics*, *22*, 97–117. doi: 10.1093/jos/ffh018

Gierasimczuk, N., van der Maas, H. L. J., & Raijmakers, M. E. J. (2013). An Analytic Tableaux Model for Deductive Mastermind Empirically Tested with a Massively Used Online Learning System. *Journal of Logic, Language and Information*, *22*, 297–314. doi: 10.1007/s10849-013-9177-5

Gil, D. (1993). Nominal and Verbal Quantification. *Sprachtypologie und Universalienforschung*, *46*, 275–317. doi: 10.1524/stuf.1993.46.14.275

Glennan, S. (2002). Rethinking Mechanistic Explanation. *Philosophy of Science*, *69*, S342–S353. doi: 10.1086/341857

Glennan, S. (2017). *The New Mechanical Philosophy.* Oxford: Oxford University Press.

Goodman, N. (1977). *The Structure of Appearance* (3rd ed.). Dordrecht: D. Reidel Publishing Company. doi: 10.1007/978-94-010-1184-6

Graf, T. (2012). Locality and the Complexity of Minimalist Derivation Tree Languages. In P. de Groot & M.-J. Nederhof (Eds.), *Formal Grammar 2010/2011* (Vol. 7395, pp. 208–227). Heidelberg: Springer. doi: 10.1007/978-3-642-32024-8_14

Graf, T. (2017). Derivations as Representations: News from the Computational Frontier. *Wiener Linguistische Gazette*, *82*, 61–69.

Graf, T. (2019). A Subregular Bound on the Complexity of Lexical Quantifiers. In J. J. Schlöder, D. McHugh, & F. Roelofsen (Eds.), *Proceedings of the 22nd Amsterdam Colloquium* (pp. 455–464).

Graham, S. M., Joshi, A., & Pizlo, Z. (2000). The traveling salesman problem: A hierarchical model. *Memory & Cognition*, *28*, 1191–1204. doi: 10.3758/BF03211820

Hackl, M. (2000). *Comparative determiners* (Unpublished doctoral dissertation). MIT, Cambridge, MA.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: most versus more than half. *Natural Language Semantics*, *17*, 63–98.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science*, *304*, 438–441. doi: 10.1126/science.1095455

Heim, I. (1999). *Notes on superlatives.* Retrieved from `https://semanticsarchive.net/Archive/TI1MTlhZ/Superlative.pdf` (Ms)

Heim, I., & Kratzer, A. (1998). *Semantics in Generative Grammar.* Malden, MA: Blackwell Publishers Inc.

Heinz, J. (2016). Computational theories of learning and developmental psycholinguistics. In J. Lidz, W. Snyder, & J. Pater (Eds.), *The oxford handbook of developmental linguistics* (pp. 633–663). Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780199601264.013.27

Heinz, J. (2018). The computational nature of phonological generalizations. In L. Hyman & F. Plank (Eds.), *Phonological Typology* (pp. 126–195). Berlin: De Gruyter Mouton.

Heinz, J., & Rawal, H. G., Chetan Tanner. (2011). Tier-based Strictly Local Constraints for Phonology. In *Proceedings of the 49th annual meeting of the association for computational linguistics* (pp. 58–64). Portland, Oregon: Association for Computational Linguistics.

Higginbotham, J., & May, R. (1981). Questions, quantifiers and crossing. *The*

*Linguistic Review*, *1*, 41–80. doi: 10.1515/tlir.1981.1.1.41

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages, and computation.* Reading, Mass: Addison-Wesley.

Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English* (Unpublished doctoral dissertation). UCLA, Los Angeles, CA.

Hubbard, R. J., Rommers, J., Jacobs, C. L., & Federmeier, K. D. (2019). Downstream Behavioral and Electrophysiological Consequences of Word Prediction on Recognition Memory. *Frontiers in Human Neuroscience*, *13*, 291. doi: 10.3389/fnhum.2019.00291

Hunter, T., & Lidz, J. (2013). Conservativity and Learnability of Determiners. *Journal of Semantics*, *30*, 315–334. doi: 10.1093/jos/ffs014

Hunter, T., Lidz, J., Odic, D., & Wellwood, A. (2017). On how verification tasks are related to verification procedures: a reply to Kotek et al. *Natural Language Semantics*, *25*, 91–107. doi: 10.1007/s11050-016-9130-7

Isaac, A. M. C., 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* (pp. 787–824). Cham: Springer. doi: 10.1007/978-3-319-06025-5_30

Israel, M. (1996). Polarity sensitivity as lexical semantics. *Linguistics and Philosophy*, *19*, 619–666. doi: 10.1007/BF00632710

Israel, M. (2001). Minimizers, Maximizers and the Rhetoric of Scalar Reasoning . *Journal of Semantics*, *18*, 297–331. doi: 10.1093/jos/18.4.297

Jackendoff, R. (2002). *Foundations of Language : Brain, Meaning, Grammar, Evolution.* Oxford: Oxford University Press.

Jacob, S. N., & Nieder, A. (2009, 4). Notation-independent representation of fractions in the human parietal cortex. *Journal of Neuroscience*, *29*(14), 4652–4657. doi: 10.1523/JNEUROSCI.0651-09.2009

Johnsen, L. G. (1987). There Sentences and Generalized Quantifiers. In P. Gärdenfors (Ed.), *Generalized Quantifiers: Linguistic and Logical Approaches* (pp. 93–107). Dordrecht: D. Reidel Publishing Company. doi: 10.1007/978-94-009-3381-1

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* Cambridge: Cambridge University Press.

Just, M. A. (1974). Comprehending quantified sentences: The relation between sentence-picture and semantic memory verification. *Cognitive Psychology*, *6*, 216–236. doi: 10.1016/0010-0285(74)90011-5

Just, M. A., & Carpenter, P. A. (1971). Comprehension of Negation with Quantification. *Journal of Verbal Learning and Verbal Behavior*, *10*, 244–253.

Kanazawa, M. (2013). Monadic Quantifiers Recognized by Deterministic Pushdown Automata. In M. Aloni, M. Franke, & F. Roelofsen (Eds.), *Proceedings of the 19th amsterdam colloquium* (pp. 139–146). Retrieved from `https://archive.illc.uva.nl/AC/AC2013/uploaded_files/\inlineitem/18_Kanazawa.pdf`

Kaplan, R. M., & Kay, M. (1994). Regular models of phonological rule systems. *Computational Linguistics*, *20*, 331–378.

Kappenman, E. S., & Luck, S. J. (2012). ERP Components: The Ups and Downs of Brainware Recordings. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford Handbook of Event-Related Potentials.* Oxford: Oxford University Press.

Katsos, N., & Cummins, C. (2010). Pragmatics: From Theory to Experiment and Back Again. *Language and Linguistics Compass*, *4*, 282–295. doi: 10.1111/j.1749-818X.2010.00203.x

Keenan, E. L. (1987). Unreducible n-ary Quantifiers in Natural Language. In P. Gärdenfors (Ed.), *Generalized quantifiers: Linguistic and logical approaches* (pp. 109–150). Dordrecht: D. Reidel Publishing Company. doi: 10.1007/978-94-009-3381-1

Keenan, E. L. (1992). Beyond the Frege Boundary. *Linguistics and Philosophy*, *15*, 199–221.

Keenan, E. L. (2002). Some properties of natural language quantifiers: Generalized quantifier theory. *Linguistics and Philosophy*, *25*, 627–654. doi: 10.1023/A:1020803514176

Keenan, E. L., & Moss, L. S. (1985). Generalized Quantifiers and the Expressive Power of Natural Language. In J. van Benthem & A. ter Meulen (Eds.), *Generalized quantifiers in natural language* (pp. 73–124). Dordrecht: Foris.

Keenan, E. L., & Paperno, D. (2017). Overview. In D. Paperno & E. L. Keenan (Eds.), *Handbook of quantifiers in natural language: Volume ii* (pp. 995–1004). Cham: Springer. doi: 10.1007/978-3-319-44330-0_20

Keenan, E. L., & Stavi, J. (1986). A Semantic Characterization Of Natural Language Determiners. *Linguistics and Philosophy*, *9*, 253-326.

Khemlani, S., & Johnson-Laird, P. N. (2022). Reasoning about properties: A computational theory. *Psychological Review*, *129*, 289–312. doi: 10.1037/rev0000240

King, J. W., & Kutas, M. (1995). Who Did What and When? Using Word- and Clause-Level ERPs to Monitor Working Memory Usage in Reading. *Journal of Cognitive Neuroscience*, *7*, 376–395. doi: 10.1162/jocn.1995.7.3.376

Kleene, S. C. (1951, 1). *Representation of Events in Nerve Nets and Finite Automata* (Tech. Rep. No. RM-704). U.S. Air Force / RAND Corporation.

Kluender, J. W., & Kutas, M. (1995). Bridging the Gap: Evidence from ERPs on the Processing of Unbounded Dependencies. *Journal of Cognitive Neuroscience*, *5*, 196–214.

Knoeferle, P., Urbach, T. P., & Kutas, M. (2011). Comprehending how visual context influences incremental sentence processing: Insights from ERPs and picture-sentence verification. *Psychophysiology*, *48*, 495–506. doi: 10.1111/j.1469-8986.2010.01080.x

Knowlton, T., Hunter, T., Odic, D., Wellwood, A., Halberda, J., Pietroski, P., & Lidz, J. (2021). Linguistic meanings as cognitive instructions. *Annals of the New York Academy of Sciences*. doi: 10.1111/nyas.14618

Knowlton, T., Pietroski, P., Halberda, J., & Lidz, J. (2022). The mental representation of universal quantifiers. *Linguistics and Philosophy*, *45*, 911–941. doi: 10.1007/s10988-021-09337-8

Knowlton, T., Trueswell, J., & Papafragou, A. (2022). A Mentalistic Semantics Explains "Each" and "Every" Quantifier Use. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th annual conference of the cognitive science society.* Retrieved from https://escholarship.org/uc/item/5qt582m2

Kusak, G., Grune, K., Hagendorf, H., & Metz, A.-M. (2000). Updating of working memory in a running memory task: an event-related potential study. *International Journal of Psychophysiology*, *39*, 51–65. doi: 10.1016/S0167-8760(00)00116-1

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes*, *8*, 533–572. doi: 10.1080/01690969308407587

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*, 621-647. doi: 10.1146/annurev.psych.093008.131123

Ladusaw, W. A. (1979). *Polarity Sensitivity as Inherent Scope Relations* (Unpublished doctoral dissertation). The University of Texas at Austin.

Langendoen, D. T. (1981). The Generative Capacity of Word-Formation Components. *Linguistic Inquiry*, *12*, 320–322.

Laszlo, S., & Federmeier, K. D. (2014). Never seem to find the time: evaluating the physiological time course of visual word recognition with regression analysis of single-item event-related potentials. *Language, Cognition and Neuroscience*, *29*, 642–661. doi: 10.1080/01690965.2013.866259

Lau, E., Almeida, D., Hines, P. C., & David, P. (2009). A lexical basis for N400 context effects: Evidence from MEG. *Brain and Language*, *111*, 161–172. doi: 10.1016/j.bandl.2009.08.007

Lau, E., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:

(de)constructing the N400. *Nature Review Neuroscience*, *9*, 920–933. doi: 10.1038/nrn2532

Lefebvre, C. D., Marchand, Y., Eskes, G. A., & Connolly, J. F. (2005). Assessment of working memory abilities using an event-related brain potential (ERP)-compatible digit span backward task. *Clinical Neurophysiology*, *116*, 1665–1680. doi: 10.1016/j.clinph.2005.03.015

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Lewis, D. (1986). Causal Explanation. In D. Lewis (Ed.), *Philosophical Papers vol. II* (pp. 214–240). Oxford: Oxford University Press.

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*, 227–256. doi: 10.1007/s11050-010-9062-6

Lindquist, M. A., Loh, J. M., Atlas, L. Y., & Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *NeuroImage*, *45*, S187–S198. doi: 10.1016/j.neuroimage.2008.10.065

Lindstrøm, P. (1966). First Order Predicate Logic with Generalized Quantifiers. *Theoria*, *32*, 186–195. doi: 10.1111/j.1755-2567.1966.tb00600.x

Lipton, P. (1991). Contrastive Explanation and Causal Triangulation. *Philosophy of Science*, *58*, 687–697. Retrieved from https://www.jstor.org/stable/188488

Lipton, P. (2004). *Inference to the Best Explanation* (2nd ed.). London: Routledge.

Lisman, J. E., & Jensen, O. (2013). The Theta-Gamma Neural Code. *Neuron*, *77*, 1002–1016. doi: 10.1016/j.neuron.2013.03.007

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique* (2nd ed.). Cambridge, MA: The MIT Press.

Luck, S. J., & Kappenman, E. S. (Eds.). (2012). *The Oxford Handbook of Event-Related Potentials*. Oxford: Oxford University Press.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*, 1–25. Retrieved from https://www.jstor.org/stable/188611

Marchand, Y., Lefebvre, C. D., & Connolly, J. F. (2006). Correlating digit span performance and event-related potentials to assess working memory. *International Journal of Psychophysiology*, *62*, 280–289. doi: 10.1016/j.ijpsycho.2006.05.007

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*, 177–190. doi:

10.1016/j.jneumeth.2007.03.024

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: W. H. Freeman.

Matthewson, L. (2001). Quantification and the Nature of Crosslinguistic Variation. *Natural Language Semantics*, *9*, 145–189. doi: 10.1023/A:1012492911285

McCawley, J. D. (1995). Generative Semantics. In E. F. K. Koerner & R. E. Asher (Eds.), *Concise History of the Language Sciences: From the Sumerians to the Cognitivists* (pp. 343–348). Oxford: Pergamon. doi: 10.1016/B978-0-08-042580-1.50057-X

McEvoy, L. K., Smith, M. E., & Gervins, A. (1998). Dynamic cortical networks of verbal and spatial working memory: Effects of memory load and task practice. *Cerebral Cortex*, *8*, 563–574. doi: 10.1093/cercor/8.7.563

McMillan, C. T., Clark, R., Moore, P., Devita, C., & Grossman, 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, 12). Quantifier comprehension in corticobasal degeneration. *Brain and Cognition*, *62*(3), 250–260. doi: 10.1016/j.bandc.2006.06.005

McNaughton, R., & Papert, S. A. (1971). *Counter-Free Automata.* Cambridge, MA: The MIT Press.

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *European Journal of Neuroscience*, *48*, 2609–2621. doi: 10.1111/ejn.13748

Mock, J., Huber, S., Bloechle, J., Bahnmueller, J., Moeller, K., & Klein, E. (2019, 7). Processing symbolic and non-symbolic proportions: Domain-specific numerical and domain-general processes in intraparietal cortex. *Brain Research*, *1714*, 133–146. doi: 10.1016/j.brainres.2019.02.029

Mock, J., Huber, S., Bloechle, J., Dietrich, J. F., Bahnmueller, J., Rennig, J., ... Moeller, K. (2018, 5). Magnitude processing of symbolic and non-symbolic proportions: An fMRI study. *Behavioral and Brain Functions*, *14*(1). doi: 10.1186/s12993-018-0141-z

Moltmann, F. (1992). Reciprocals and same/different: Towards a semantic analysis. *Linguistics and Philosophy*, *15*, 411–462.

Moltmann, F. (1995). Exception sentences and polyadic quantification. *Linguistics and Philosophy*, *18*, 223–280.

Moltmann, F. (2003). Propositional Attitudes Without Propositions. *Synthese*, *135*, 77–118. doi: 10.1023/A:1022945009188

Morgan, B., Gross, R. G., Clark, R., Dreyfuss, M., Boller, A., Camp, E., ... Grossman, M. (2011, 11). Some is not enough: Quantifier comprehension in corticobasal syndrome and behavioral variant frontotemporal dementia. *Neuropsychologia*, *49*(13), 3532–3541. doi: 10.1016/j.neuropsychologia.2011.09.005

Moschovakis, Y. N. (2006). A Logical Calculus of Meaning and Synonymy. *Linguistics and Philosophy*, *29*, 27–89. doi: 10.1007/s10988-005-6920-7

Mostowski, A. (1957). On a generalization of quantifiers . *Fundamenta Mathematicae*, *44*, 12–36.

Mostowski, M. (1991). Divisibility Quantifiers. *Bulletin of the Section of Logic*, *20*, 67–70.

Mostowski, M. (1998). Computational Semantics for Monadic Quantifiers. *Journal of Applied Non-Classical Logics*, *8*, 107–121. doi: 10.1080/11663081.1998.10510934

Müller, H. M., King, J. W., & Kutas, M. (1997). Event-related potentials elicited by spoken relative clauses. *Cognitive Brain Research*, *5*, 193–203. doi: 10.1016/S0926-6410(96)00070-5

Münte, T. F., Schiltz, K., & Kutas, M. (1998). When temporal terms belie conceptual order. *Nature*, *395*, 71–73. doi: 10.1038/25731

Muskens, R. (2005). Sense and the Computation of Reference. *Linguistics and Philosophy*, *28*, 473–504. doi: 10.1007/s10988-004-7684-1

Näätänen, R., & Kreegipuu, K. (2012). The Mismatch Negativity (MMN). In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford Handbook of Event-Related Potentials.* Oxford: Oxford University Press.

Nam, S. (2005). n-ary Quantifiers and the Expressive Power of DP-compositions. *Research on Language and Computation*, *3*, 411–428. doi: 10.1007/s11168-006-0005-9

Nieder, A., & Dehaene, S. (2009). Representation of Number in the Brain. *Annual Review of Neuroscience*, *32*, 185-208. doi: 10.1146/annurev.neuro.051508.135550

Nieuwland, M. S. (2016). Quantification, Prediction, and the Online Impact of Sentence Truth-Value: Evidence From Event-Related Potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 316–334. doi: 10.1037/xlm0000173

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., ... Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*. doi: 10.1098/rstb.2018.0522

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, *63*, 324–346. doi:

10.1016/j.jml.2010.06.005

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*, 1213–1218. doi: 10.1111/j.1467-9280.2008.02226.x

Nieuwland, M. S., & Martin, A. R. (2012). If the real world were irrelevant, so to speak: The role of propositional truth-value in counterfactual sentence comprehension. *Cognition*, *122*, 102–109. doi: 10.1016/j.cognition.2011.09.001

Nieuwland, M. S., & van Berkum, J. J. A. (2006). When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience*, *18*, 1098–1111. doi: 10.1162/jocn.2006.18.7.1098

Niyogi, P. (2006). *The Computational Nature of Language Learning and Evolution.* Cambridge, MA: The MIT Press.

Noveck, I. A., & Reboul, A. (2008). Experimental Pragmatics: a Gricean turn in the study of language. *Trends in Cognitive Sciences*, *12*, 425–431. doi: 10.1016/j.tics.2008.07.009

Odic, D., & Starr, A. (2018). An Introduction to the Approximate Number System. *Child Development Perspectives*, *12*, 223–229. doi: 10.1111/cdep.12288

Olm, C. A., McMillan, C. T., Spotorno, N., Clark, R., & Grossman, M. (2014, 8). The relative contributions of frontal and parietal cortex for generalized quantifier comprehension. *Frontiers in Human Neuroscience*, *8*(AUG). doi: 10.3389/fnhum.2014.00610

O'Rourke, P. L., & Van Petten, C. (2011). Morphological agreement at a distance: Dissociation between early and late components of the event-related brain potential. *Brain Research*, *1392*, 62–79. doi: 10.1016/j.brainres.2011.03.071

Pagin, P. (2012). Communication and the Complexity of Semantics. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford Handbook of Compositionality* (pp. 510–529). Oxford: Oxford University Press.

Partee, B. H. (1989). Many quantifiers. In J. Powers & K. de Jong (Eds.), *Escol 89: Proceedings of the eastern states conference on linguistics* (pp. 383–402). Columbus, OH: Department of Linguistics, Ohio State University.

Partee, B. H. (1995). Quantificational Structures and Compositionality. In E. Bach, E. Jelinek, A. Kratzer, & B. H. Partee (Eds.), *Quantification in Natural Languages* (p. 541-601). Dordrecht: Kluwer Academic Publishers. doi: 10.1007/978-94-017-2817-1

Pelosi, L., Hayward, M., & Blumhardt, L. D. (1995). Is "memory-scanning" time in the Sternberg paradigm reflected in the latency of event-related potentials? *Electroencephalography and Clinical Neurophysiology*, *96*, 44–55. doi: 10.1016/0013-4694(94)00163-F

Pelosi, L., Hayward, M., & Blumhardt, L. D. (1998). Which event-related potentials reflect memory processing in a digit-probe identification task? *Cognitive Brain Research*, *6*, 205–218. doi: 10.1016/S0926-6410(97)00032-3

Pelosi, L., Holly, M., Slade, T., Hayward, M., Barrett, G., & Blumhardt, L. D. (1992). Wave form variations in auditory event-related potentials evoked by a memory-scanning task and their relationship with tests of intellectual function. *Electroencephalography and Clinical Neurophysiology*, *84*, 344–352. doi: 10.1016/0168-5597(92)90087-R

Peters, S., & Westerståhl, D. (2006). *Quantifiers in Language and Logic.* Oxford: Oxford University Press.

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The Meaning of 'Most': Semantics, Numerosity and Psychology. *Mind and Language*, *24*, 554–585. doi: 10.1111/j.1468-0017.2009.01374.x

Pietroski, P., Lidz, J., Hunter, T., Odic, D., & Halberda, J. (2011). Seeing what you mean, mostly. In J. Runner (Ed.), *Experiments at the interfaces* (Vol. 37, pp. 181–217). Leiden: Brill. doi: 10.1108/S0092-4563(2011)0000037010

Plöchl, M., Ossandón, J. P., & König, P. (2012). Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in Human Neuroscience*, *6*. doi: 10.3389/fnhum.2012.00278

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*, 59–63. doi: 10.1016/j.tics.2005.12.004

Polich, J. (2012). Neuropsychology of the P300. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford Handbook of Event-Related Potentials.* Oxford: Oxford University Press.

Polimeni, J. R., & Lewis, L. D. (2021). Imaging faster neural dynamics with fast fMRI: A need for updated models of the hemodynamic response. *Progress in Neurobiology*, *207*, 102174. doi: 10.1016/j.pneurobio.2021.102174

Pratt-Hartmann, I. (2004). Fragments of Language. *Journal of Logic, Language and Information*, *13*, 207–223. doi: 10.1023/B:JLLI.0000024735.97006.5a

Pustejovsky, J. (1995). *The Generative Lexicon.* Cambridge, MA: The MIT Press.

Ramotowska, S., Steinert-Threlkeld, S., van Maanen, L., & Szymanik, J. (2020). Most, but not more than half, is proportion-dependent and sensitive to individual differences. In M. Franke, N. Kompa, M. Liu, J. L. Mueller, & J. Schwab (Eds.), *Proceedings of sinn und bedeutung 24* (Vol. 2, pp. 165–182). Osnabrück University.

Ratcliff, R., Sederberg, P. B., Smith, T. A., & Childers, R. (2016). A single trial analysis of EEG in recognition memory: Tracking the neural correlates of memory strength. *Neuropsychologia*, *93*, 128–141. doi: 10.1016/j.neuropsychologia.2016.09.026

Rett, J. (2015). *The semantics of evaluativity.* Oxford: Oxford University Press.

Rett, J. (2018). The semantics ofmany, much, few, and little. *Language and Linguistics Compass*, *12*. doi: 0.1111/lnc3.12269

Ristad, E. S. (1993). *The Language Complexity Game.* Cambridge, MA: The MIT Press.

Rogers, J., Heinz, J., Fero, M., Hurst, J., Lambert, D., & Wibel, S. (2013). Cognitive and Sub-regular Complexity. In G. Morrill & M.-J. Nederhof (Eds.), *Formal grammar* (Vol. 8036, pp. 90–108). Dordrecht: Springer. doi: 10.1007/978-3-642-39998-5_6

Romero, M. (1998). *Focus and reconstruction effects in wh-phrases* (Unpublished doctoral dissertation). University of Massachusetts, Amherst, MA.

Romero, M. (2015). The Conservativity of many. In T. Brochhagen, F. Roelofsen, & N. Theiler (Eds.), *Proceedings of the 20th Amsterdam Colloquium* (pp. 20–29). Amsterdam: Institute for Logic, Language and Computation.

Rommers, J., & Federmeier, K. D. (2018). Electrophysiological Methods. In A. M. B. de Groot & P. Hagoort (Eds.), *Research methods in psycholinguistics and the neurobiology of language: A practical guide* (pp. 247–265). Oxford: John Wiley & Sons.

Romoli, J. (2015). Toward a structural account of Conservativity. *Semantics-Syntax Interface*, *2*, 28–57.

Rösler, F., Heil, M., & Röder, B. (1997). Slow negative brain potentials as reflections of specific modular resources of cognition. *Biological Psychology*, *45*, 109–141. doi: 10.1016/S0301-0511(96)05225-8

Ruchkin, D. S., Johnson, R., Canoune, H., & Ritter, W. (1990). Short-term memory storage and retention: an event-related brain potential study. *Electroencephalography and Clinical Neurophysiology*, *76*, 419–439. doi: 10.1016/0013-4694(90)90096-3

Ruchkin, D. S., Johnson, R., Grafman, J., Canoune, H., & Ritter, W. (1992). Distinctions and similarities among working memory processes: an event-related potential study. *Cognitive Brain Research*, *1*, 53–66. doi: 10.1016/0926-6410(92)90005-C

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high- and low-frequency words. *Memory & Cognition*, *18*, 367–379.

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *TRENDS in Cognitive Sciences*, *11*, 251-257. doi: 10.1016/j.tics.2007.04.004

Rugg, M. D., Mark, R. E., Walla, P., Schloerscheidt, A. M., Birch, C. S., & Allan, K. (1998). Dissociation of the neural correlates of implicit and explicit memory. *Nature*, *392*, 595-–598. doi: 10.1038/33396

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World.*

Princeton, NJ: Princeton University Press.

Salmon, W. C. (1989). *Four Decades of Scientific Explanation.* Minneapolis, MN: University of Minnesota Press.

Sassenhagen, J., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2014). The P600-as-P3 hypothesis revisited: single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain and Language*, *137*, 29–39. doi: 10.1016/j.bandl.2014.07.010

Schlotterbeck, F., Ramotowska, S., van Maanen, L., & Szymanik, J. (2020). Representational complexity and pragmatics cause the monotonicity effect. In S. Denison, M. Mack, Y. Xu, & B. Armstrong (Eds.), *Proceedings of the 42nd annual meeting of the cognitive science society* (pp. 3398–3404). Cognitive Science Society.

Sharvit, Y., & Stateva, P. (2002). Superlative expressions, context, and focus. *Linguistics and Philosophy*, *25*, 453–504. doi: 10.1023/A:1020875809794

Shieber, S. M. (1985). Evidence against the Context-Freeness of Natural Language. *Linguistics and Philosophy*, *8*, 333–345. doi: 10.1007/BF00630917

Shikhare, S., Heim, S., Klein, E., Huber, S., & Willmes, K. (2015). Processing of Numerical and Proportional Quantifiers. *Cognitive Science*, *39*, 1504–1536. doi: 10.1111/cogs.12219

Skagenholt, M., Träff, U., Västfjäll, D., & Skagerlund, K. (2018, 6). Examining the triple code model in numerical cognition: An fmri study. *PLoS ONE*, *13*(6). doi: 10.1371/journal.pone.0199247

Steinert-Threlkeld, S., & Icard III, T. F. (2013). Iterated semantic automata. *Linguistics and Philosophy*, *36*, 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. Roelofsen, & N. Theiler (Eds.), *Proceedings of the 20th amsterdam colloquium* (pp. 368–377).

Steinert-Threlkeld, S., & Szymanik, J. (2019). Learnability and semantic universals. *Semantics and Pragmatics*, *12*. doi: 10.3765/sp.12.4

Suppes, P. (1982). Variable-free semantics with remarks on procedural extensions. In T. W. Simon & R. J. Scholes (Eds.), *Language, mind, and brain* (pp. 21–34). Hillsdale: Erlbaum.

Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2012). Language-Related ERP Components. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford Handbook of Event-Related Potentials.* Oxford: Oxford University Press.

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. (2010). Computational complexity of polyadic lifts of generalized quantifiers in natural language. *Linguistics and Philosophy*, *33*, 215–250. doi: 10.1007/s10988-010-9076-z

Szymanik, J. (2016). *Quantifiers and Cognition: Logical and Computational Perspectives*. Cham: Springer. doi: 10.1007/978-3-319-28749-2

Szymanik, J., Meijering, B., & Verbrugge, R. (2013). Using intrinsic complexity of turn-taking games to predict participants' reaction times. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 1426–1431). Austin, TX: Cognitive Science Society.

Szymanik, J., Steinert-Threlkeld, S., Zajenkowski, M., & Icard III, T. F. (2013). Automata and Complexity in Multiple-Quantifier Sentence Verification. In R. L. West & T. C. Stewart (Eds.), *Proceedings of the 12th International Conference on Cognitive Modeling* (p. 239-245). Ottawa: Carleton University.

Szymanik, J., & Thorne, C. (2017, 3). Exploring the relation between semantic complexity and quantifier distribution in large corpora. *Language Sciences*, *60*, 80–93. doi: 10.1016/j.langsci.2017.01.006

Szymanik, J., & Zajenkowski, M. (2009). Improving methodology of quantifier comprehension experiments. *Neuropsychologia*, *47*(12), 2682–2683. doi: 10.1016/j.neuropsychologia.2009.04.004

Szymanik, J., & Zajenkowski, M. (2010a, 4). Comprehension of simple quantifiers: Empirical evaluation of a computational model. *Cognitive Science*, *34*(3), 521–532. doi: 10.1111/j.1551-6709.2009.01078.x

Szymanik, J., & Zajenkowski, M. (2010b). Quantifiers and Working Memory. In M. Aloni, H. Bastiaanse, T. de Jager, P. van Ormondt, & K. Schulz (Eds.), *Amsterdam colloquium 2009* (Vol. 25, pp. 456–464). Berlin Heidelberg: Springer Verlag.

Szymanik, J., & Zajenkowski, M. (2011, 12). Contribution of working memory in parity and proportional judgments. *Belgian Journal of Linguistics*, *25*, 176–194. doi: 10.1075/bjl.25.08szy

Talmina, N., Kochari, A., & Szymanik, J. (2017). Quantifiers and verification strategies: connecting the dots. In A. Cremers, T. van Gessel, & F. Roelofsen (Eds.), *Proceedings of the 21st amsterdam colloquium* (pp. 465–473).

ter Meulen, A. (1991). English aspectual verbs as generalized quantifiers. In A. L. Halpern (Ed.), *Proceedings of the 9th west coast conference on formal linguistics* (pp. 347–360).

Tessler, M. H., Tenenbaum, J. B., & Goodman, N. D. (2022). Logic, Probability, and Pragmatics in Syllogistic Reasoning. *Topic in Cognitive Science*. doi: 10.1111/tops.12593

Thorne, C. (2012). Studying the Distribution of Fragments of English Using Deep Semantic Annotation. In H. Bunt (Ed.), *Proceedings of the 8th workshop in semantic annotation (ISA 8).*

Tichý, P. (1969). Intension in terms of Turing machines. *Studia Logica*, *24*, 7–21. doi: 10.1007/BF02134290

Tomaszewicz, B. (2011). Verification Strategies for Two Majority Quantifiers in Polish. In I. Reich, E. Horch, & D. Pauly (Eds.), *Proceedings of sinn und bedeutung 15* (pp. 597–612). Retrieved from `https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/402`

Tomaszewicz-Özakın, B. (2020). The Semantics of the Superlative Quantifier -Est. In P. Hallman (Ed.), *Interactions of Degree and Quantification* (pp. 79–120). Leiden: Brill.

Troiani, V., Peelle, J. E., McMillan, C., Clark, R., & Grossman, M. (2009a). Magnitude and parity as complementary attributes of quantifier statements. *Neuropsychologia*, *47*(12), 2684–2685. doi: 10.1016/j.neuropsychologia.2009.04.025

Troiani, V., Peelle, J. E., McMillan, C., Clark, R., & Grossman, M. (2009b). Magnitude and parity as complementary attributes of quantifier statements. *Neuropsychologia*, *47*(12), 2684–2685. doi: 10.1016/j.neuropsychologia.2009.04.025

Tsogli, V., Jentschke, S., & Koelsch, S. (2022). Unpredictability of the "when" influences prediction error processing of the "what" and "where". *PLOS One*, *17*. doi: 10.1371/journal.pone.0263373

Urbach, T. P., DeLong, K. A., & Kutas, M. (2015, 8). 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, 8). 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. (1984). Questions about Quantifiers. *The Journal of Symbolic Logic*, *49*, 443-466. doi: 10.2307/2274176

van Benthem, J. (1986a). Determiners. In *Essays in Logical Semantics* (pp. 3–24). Dordrecht: D. Reidel Publishing Company.

van Benthem, J. (1986b). *Essays in Logical Semantics.* Dordrecht: D. Reidel Publishing Company.

van Benthem, J. (1986c). Semantic Automata. In *Essays in Logical Semantics* (pp. 151–176). Dordrecht: D. Reidel Publishing Company.

van Benthem, J. (1989). Polyadic Quantifiers. *Linguistics and Philosophy*, *12*, 437–464.

van Benthem, J., & ter Meulen, A. (Eds.). (1985). *Generalized Quantifiers in Natural Language*. Dordrecht: Foris.

van Berkum, J. J. A., Brown, C. M., & Hagoort, P. (1999). Early Referential Context Effects in Sentence Processing: Evidence from Event-Related Brain Potentials. *Journal of Memory and Language*, *41*, 147–182. doi: 10.1006/jmla.1999.2641

van Berkum, J. J. A., Brown, C. M., Hagoort, P., & Zwitserlood, P. (2003). Event-related brain potentials reflect discourse-referential ambiguity in spoken language comprehension. *Psychophysiology*, *40*, 235–248. doi: 10.1111/1469-8986.00025

van de Pol, I., Lodder, P., van Maanen, L., Steinert-Threlkeld, S., & Szymanik, J. (2021). Quantifiers satisfying semantic universals are simpler. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd annual meeting of the cognitive science society.* Retrieved from `https://escholarship.org/uc/item/1vm445rp`

van de Pol, I., Steinert-Threlkeld, S., & Szymanik, J. (2019). Complexity and learnability in the explanation of semantic universals of quantifiers. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual conference of the cognitive science society* (pp. 3015–3021). Montreal, QB: Cognitive Science Society.

van der Hoek, W., & de Rijke, M. (1993). Generalized quantifiers and modal logic. *Journal of Logic, Language and Information*, *2*, 19–58. doi: 10.1007/BF01051767

van Fraassen, B. C. (1977). The Pragmatics of Explanation. *American Philosophical Quarterly*, *14*, 143–150. Retrieved from `https://www.jstor.org/stable/20009661`

van Fraassen, B. C. (1980). *The Scientific Image.* Oxford: Clarendon Press.

van Lambalgen, M., & Hamm, F. (2005). *The Proper Treatment of Events.* Malden: Blackwell.

Van Petten, C. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition*, *18*, 380–393. doi: 10.3758/BF03197127

Van Petten, C. (2014). Examining the N400 semantic context effect item-by-item: Relationship to corpus-based measures of word co-occurrence. *International Journal of Psychophysiology*, *94*, 407–419. doi: 10.1016/j.ijpsycho.2014.10.012

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time Course of Word Identification and Semantic Integration in Spoken Language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 394–417. doi: 10.1037/0278-7393.25.2.394

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*, 176–190. doi: 10.1016/j.ijpsycho.2011.09.015

van Rooij, I. (2008). The Tractable Cognition Thesis. *Cognitive Science*, *32*, 939–984. doi: 10.1080/03640210801897856

van Rooij, I., & Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychological Inquiry*, *31*(4), 321–325.

van Rooij, I., & Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science*. doi: 10.1177/1745691620970604

van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and Intractability: A Guide to Classical and Parameterized Complexity Analysis.* Cambridge: Cambridge Univeristy Press. doi: 10.1017/9781107358331

van Rooij, I., Schactman, A., Kadlec, H., & Stege, U. (2006). Perceptual or Analytical Processing? Evidence from Children's and Adult's Performance on the Euclidean Traveling Salesperson Problem. *Journal of Problem Solving*, *1*, 44–73. doi: 10.7771/1932-6246.1006

van Rooij, I., & Wareham, T. (2012). Intractability and approximation of optimization theories of cognition. *Journal of Mathematical Psychology*, *56*, 232–247. doi: 10.1016/j.jmp.2012.05.002

van Rooij, I., Wright, C. D., & Wareham, T. (2012). Intractability and the use of heuristics in psychological explanations. *Synthese*, *187*, 471–487. doi: 0.1007/s11229-010-9847-7

Vissers, C. T. W. M., Kolk, H. K. J., van de Meerendonk, N., & Chwilla, D. J. (2008). Monitoring in language perception: Evidence from ERPs in a picture–sentence matching task. *Neuropsychologia*, *46*, 967–982. doi: 10.1016/j.neuropsychologia.2007.11.027

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*, 748–751. doi: 10.1038/nature02447

von Fintel, K. (1997). Bare Plurals, Bare Conditionals, and Only. *Journal of Semantics*, *14*, 1–56. doi: 10.1093/jos/14.1.1

von Fintel, K., & Keenan, E. L. (2018). Determiners, Conservativity, Witnesses. *Journal of Semantics*, *35*, 207–217. doi: 10.1093/jos/ffx018

von Fintel, K., & Matthewson, L. (2008). Universals in Semantics. *The Linguistic Review*, *25*, 139–201. doi: 10.1515/TLIR.2008.004

Vos, S. H., Gunter, T. C., Kolk, H. H. J., & Mulder, G. (2001). Working memory constraints on syntactic processing: An electrophysiological investigation. *Psychophysiology*, *38*, 41–63. doi: 10.1111/1469-8986.3810041

Wassenaar, M., & Hagoort, P. (2007). Thematic role assignment in patients with Broca's aphasia: Sentence–picture matching electrified. *Neuropsychologia*, *45*, 716-740. doi: 10.1016/j.neuropsychologia.2006.08.016

Weber, E., Van Bouwel, J., & De Vreese, L. (2013). *Scientific Explanation*. Dordrecht: Springer. doi: 10.1007/978-94-007-6446-0

Westerståhl, D. (1985). Logical constants in quantifier languages. *Linguistics and Philosophy*, *8*, 387–413. doi: 10.1007/BF00637410

Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford: Oxford University Press.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels in explanation. *Biology and Philosophy*, *25*, 287–318. doi: 10.1007/s10539-010-9200-z

Yang, H., Laforge, G., Stojanoski, B., Nichols, E. S., McRae, K., & Köhler, S. (2019). Late positive complex in event-related potentials tracks memory signals when they are decision relevant. *Scientific Reports*, *9*, 9469. doi: 10.1038/s41598-019-45880-y

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, *46*, 441–517. doi: 10.1006/jmla.2002.2864

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*, 595–600. doi: 10.1016/j.jcomdis.2011.07.005

Zajenkowski, M., & Szymanik, J. (2013, 9). MOST intelligent people are accurate and SOME fast people are intelligent. Intelligence, working memory, and semantic processing of quantifiers from a computational perspective. *Intelligence*, *41*(5), 456–466. doi: 10.1016/j.intell.2013.06.020

Zajenkowski, M., Szymanik, J., & Garraffa, M. (2014, 11). Working Memory Mechanism in Proportional Quantifier Verification. *Journal of Psycholinguistic Research*, *43*(6), 839–853. doi: 10.1007/s10936-013-9281-3

Zhao, B., van de Pol, I., Raijmakers, M. E. J., & Szymanik, J. (2018). Predicting Cognitive Difficulty of the Deductive Mastermind Game with Dynamic Epistemic Logic Models. In T. Rogers, M. Rau, X. Zhu, & C. Kalish (Eds.), *Proceedings of the 40th annual meeting of the cognitive science society* (pp. 2789–2794). Austin, TX: Cognitive Science Society.

Zuber, R. (2004). A class of non-conservative determiners in Polish. *LingvisticæInvestigationes*, *27*, 147–165. doi: 10.1075/li.27.1.07zub

Zuber, R., & Keenan, E. L. (2019). A note on conservativity. *Journal of Semantics*, *35*, 573–582. doi: 10.1093/jos/ffz007

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*, 162–185. doi: 10.1037/0033-2909.123.2.162

# Papers

# Paper 1

Computational complexity explains neural differences in quantifier verification

Cognition 223 (2022) 105013

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

# Computational complexity explains neural differences in quantifier verification

Heming Strømholt Bremnes [a,*], Jakub Szymanik [b], Giosuè Baggio [a]

[a] *Language Acquisition and Language Processing Lab, Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway*
[b] *Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands*

ARTICLE INFO

*Keywords:*
Natural language quantifiers
Computational complexity
Semantic automata
Picture-sentence verification
Event-related potentials

ABSTRACT

Different classes of quantifiers provably require different verification algorithms with different complexity profiles. The algorithm for proportional quantifiers, like 'most', is more complex than that for nonproportional quantifiers, like 'all' and 'three'. We tested the hypothesis that different complexity profiles affect ERP responses during sentence verification, but not during sentence comprehension. In experiment 1, participants had to determine the truth value of a sentence relative to a previously presented array of geometric objects. We observed a sentence-final negative effect of truth value, modulated by quantifier class. Proportional quantifiers elicited a sentence-internal positivity compared to nonproportional quantifiers, in line with their different verification profiles. In experiment 2, the same stimuli were shown, followed by comprehension questions instead of verification. ERP responses specific to proportional quantifiers disappeared in experiment 2, suggesting that they are only evoked in a verification task and thus reflect the verification procedure itself. Our findings demonstrate that algorithmic aspects of human language processing are subjected to the same formal constraints applicable to abstract machines.

## 1. Introduction

Quantifiers are linguistic expressions that denote quantities and relate sets of objects. The ability to quantify is fundamental to human cognition. It is therefore not surprising that quantifiers are ubiquitous in natural languages, logic, and mathematics. Somewhat more surprisingly, given their superficial morphosyntactic diversity – ranging from simple determiners such as 'all' to multiple conjoined phrases like 'less than half and more than a third' – natural language quantifiers are remarkably invariant cross-linguistically (Bach et al., 1995; Keenan & Paperno, 2017; Matthewson, 2001) and constitute a small subset of the mathematically possible quantifiers (Barwise & Cooper, 1981; Keenan & Stavi, 1986). Furthermore, their characteristic formal properties delineate learning and processing biases in quantitative tasks for humans, non-human primates, and machine learning algorithms alike (Carcassi, Steinert-Threlkeld, & Szymanik, 2021; Chemla, Dautriche, Buccola, & El Fagot, 2019; Hunter & Lidz, 2013; Steinert-Threlkeld & Szymanik, 2020; van de Pol, Steinert-Threlkeld, & Szymanik, 2019).

For these and other reasons, quantifiers have been studied extensively in theoretical linguistics, psycholinguistics, and cognitive neuroscience. One common theme in the cognitive neuroscience literature is that quantifiers can give rise to different truth-conditions depending on the surrounding linguistic context (Freunberger & Nieuwland, 2016; Kounios & Holcomb, 1992; Nieuwland, 2016; Noveck & Posada, 2003; Urbach et al., 2015; Urbach & Kutas, 2010) or the order of the quantifiers in multiply quantified sentences (Dwivedi et al., 2010; McMillan et al., 2013). One empirical question is whether quantified sentences are verified and interpreted incrementally or whether instead their interpretation is delayed until the whole sentence has been parsed. Another question is whether incrementality interacts with negation or negative polarity more generally (Augurzky et al., 2020a; Freunberger & Nieuwland, 2016; Nieuwland, 2016; Urbach et al., 2015; Urbach & Kutas, 2010).

What unifies these studies is that they all use verification paradigms. As will be more thoroughly discussed in Section 1.1, different classes of quantifiers require distinct verification procedures, and these can in turn be classified differently in terms of their computational complexity. The aims of the present study are to explicitly manipulate quantifier class in a verification task, to demonstrate that computational complexity plays a role in determining which type of algorithm is implemented in the verification of different classes of quantifiers, and to gather initial empirical information on how quantifiers are verified by the brain.

[*] Corresponding author at: Department of Language and Literature, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway.
*E-mail address:* heming.s.bremnes@ntnu.no (H.S. Bremnes).

https://doi.org/10.1016/j.cognition.2022.105013
Received 11 October 2021; Received in revised form 29 December 2021; Accepted 4 January 2022
Available online 20 January 2022

Aside from being relevant to the processing of quantifiers specifically, the approach exemplified herein can help shed light on algorithmic aspects of semantic processing more generally – an area that hitherto has not received sufficient attention (Baggio, 2018, 2020). Arguably, in order to explain the capacity to comprehend and produce meaningful utterances, it is not enough to know what computation is being carried out and which brain areas are activated when over the course of the computation. In line with Marr's (1982) levels of analysis in cognitive science, algorithms are essential in mediating between the computational and implementational levels, since they are restricted both by the nature of the computation and by what kinds of processes can be carried out by the physical medium of the brain (Baggio, Stenning, & van Lambalgen, 2016; Baggio, van Lambalgen, & Hagoort, 2015; Embick & Poeppel, 2015; Lewis & Phillips, 2015). Regardless of the cognitive plausibility of truth functional semantics, verification is a well-defined computation, and knowing the impact of different verification procedures on sentence processing is, at a minimum, useful in disentangling effects of task from effects of representation, structure-building, prediction, and other processes.

Relatedly, there is a growing body of literature advocating so-called procedural semantics (Moschovakis, 2006; Muskens, 2005; Pietroski et al., 2009; Suppes, 1982; Szymanik, 2016; Tichý, 1969; van Benthem, 1986; van Lambalgen & Hamm, 2005), where the meaning of an expression is a set of algorithms computing its extension, which for declarative sentences amounts to a model-building or verification procedure. However, the theory we test and the task we employ here are focused on verification, not meaning representation as such. Consequently, the data cannot be used to argue for or against this philosophical position about the nature of meaning or its linguistic and computational instantiations.

### 1.1. Quantifier automata and the computational complexity of verification

Originating with van Benthem's (1986) seminal paper 'Semantic Automata', the computational properties of different quantifier expressions have been extensively studied (e.g. Kanazawa, 2013; Mostowski, 1998; Szymanik, 2016). A consequence of van Benthem's work is that proportional quantifiers – e.g., 'most', 'less than half', 'a third' etc. – are provably more computationally complex to verify than nonproportional quantifiers – expressions containing, e.g., Aristotelian quantifiers like 'all' and 'some' or numerical quantifiers like 'three' and 'five'.

Informally, verification algorithms go through the objects in the domain denoted by the quantified phrase sequentially in order to check whether the property predicated of these objects holds true. For *Aristotelian quantifiers*, this entails going through the contextually relevant objects one after the other and looking for a (counter)example of an object with(out) the predicated property; once the (counter)example is (not) found, it can be established whether the expression is true. To exemplify, when verifying a sentence like 'All the circles are red' in a domain of differently colored circles, the algorithm searches through all the circles until it finds a non-red circle, in which case the sentence is false. If a non-red circle is not found, the sentence is true. In the same vein, for *numerical quantifiers*, one counts the number of objects with the predicated property, and if one finds the number of objects required by the quantifier, the quantifier expression is true. As an illustration, consider the sentence 'Three of the circles are red' in a domain as above. For this sentence, the algorithm looks for red circles and counts until three red circles have been found. If three red circles are found, the algorithm outputs true, and if not, it outputs false. Because these algorithms only require paying attention to one type of object, either with or without the predicated property, these kinds of quantifiers can all be computed by a finite state automaton (FSA) and can equivalently be described in a regular language (Kleene, 1951).

To verify *proportional quantifiers*, by contrast, one needs to enumerate both the objects that have the predicated property and those that do not. Once one has considered and classified all the objects, one compares the number of objects in the two sets. If the ratio of objects with the predicated property to objects without it conforms to the ratio set by the quantifier, e.g., 'more than half', the expression is true. In a domain corresponding to the examples above, to verify a sentence like 'most circles are red' the algorithm must keep track of both the red circles and the non-red circles, and if the red circles outnumber the non-red circles, the algorithm outputs true; it outputs false if there are more non-red than red circles. Such verification algorithms for proportional quantifiers cannot be computed by an FSA, and instead require a push-down automaton (PDA) with a memory component where the information about both types of objects can be stored. PDAs correspond to context-free languages (Hopcroft & Ullman, 1979, p. 116), and are thus strictly more complex than regular languages – and FSAs – according to the Chomsky hierarchy (Chomsky, 1956). For a formal description and textbook explanation of the different algorithms, see Szymanik (2016, chapter 4).

### 1.2. Previous research and relevant electrophysiological effects

Previous studies have shown that computational differences between quantifiers have significant cognitive effects in terms of accuracy and reaction time in picture-sentence verification tasks (Szymanik & Zajenkowski, 2009, 2010, 2011; Zajenkowski & Szymanik, 2013; Zajenkowski et al., 2014). Furthermore, fMRI studies (McMillan et al., 2005; Olm et al., 2014) have found that (pre)frontal areas associated with working memory and executive function, notably the dorsolateral prefrontal cortex, have found an increase in BOLD responses for proportional relative to nonproportional quantifiers in the same type of task. Building on these findings, verification paradigm studies of patients with neurodegenerative diseases (McMillan et al., 2006; Morgan et al., 2011) have found that atrophy in these regions is associated with decreased performance with proportional, but not nonproportional quantifiers. Similar effects are also found in fMRI experiments in the mathematical cognition literature, where bilateral frontal activation is associated with processing of proportions both in adaptation and magnitude comparison paradigms (Jacob & Nieder, 2009; Mock et al., 2018, 2019). The same effects are found regardless of whether proportions are presented mathematically or verbally, i.e., by means of a natural language quantifier (Jacob & Nieder, 2009).

By contrast, previous electrophysiological studies of quantifiers have either considered only one class of quantifiers in each experiment (Augurzky et al., 2017; Augurzky et al., 2019; Augurzky et al., 2020a; Augurzky et al., 2020b; Kounios & Holcomb, 1992; Noveck & Posada, 2003), or have used quantifiers from different classes as polar opposites (Freunberger & Nieuwland, 2016; Nieuwland, 2016; Urbach et al., 2015; Urbach & Kutas, 2010). To our knowledge, the only exception is a small-scale study by De Santo et al. (2019), to be discussed below, that looked at differences between Aristotelian 'some' and proportional 'most'.

Additionally, few studies have looked at sentence verification in relation to a picture. Spychalska et al. (2019, 2016) were only interested in sentence final effects of implicature violations, and showed the picture mid-sentence, immediately before the final word. This modulated the N400 and post-N400 positivities. The authors were able to show that participants' pragmatic sensitivity had an effect on the evoked potential in trials where scalar implicatures were modulated. However, the design did not allow investigating incremental effects of verification that could originate at earlier points in the sentence. Hunt III et al. (2013) and Politzer-Ahles et al. (2013) were also interested in implicature violations, but presented pictures before each sentence. The former found graded N400 responses with a visual world paradigm for true, underinformative and false sentences: false sentences elicited the strongest effect compared to true, whereas underinformative fell in the middle. Politzer-Ahles et al. (2013) looked at effects on the quantifier. In a 2 × 2 design with 'some' and 'all' – where 'all' was true when 'some' was underinformative, and false when ‘some’ was strictly true – they found sustained positivities for quantificational violations with ‘all’, but sustained negativities for implicature violations with ‘some’. Augurzky et al. (2017, 2019, 2020a, 2020b) have all addressed issues of incrementality. They found that, regardless of quantifier type – Aristotelian or proportional, in nominal, e.g., ‘all the circles’, or adverbial form, e.g., ‘every day’ – the N400, and related truth value effects, are only found at the position where the sentence is disambiguated. When the presented linguistic material is compatible with the sentence being both true and false, N400 effects do not arise. The only exception to this pattern is the negative proportional quantifier ‘less than half’, for which the N400 does not arise at all (see also Nieuwland, 2016; Urbach et al., 2015; Urbach & Kutas, 2010). In these cases, they instead found an increased positivity on the quantifier, which they attributed to the semantic complexity of the negative polarity (see e.g. Deschamps et al., 2015; Just & Carpenter, 1971). In all experiments, a sustained positivity was also found after the N400 in false trials where the truth value could not be known immediately, but only when participants performed a verification task. The authors attributed this to increased attention to the picture-sentence mapping in complex contexts, and argue that it is a P600-as-P3 decision effect (Sassenhagen et al., 2014).

De Santo et al. (2019) conducted a small-scale study ($N = 8$) where they compared proportional and Aristotelian quantifiers in a picture-verification task in which participants saw an array of geometrical shapes while hearing a quantified sentence. The auditory stimuli were divided into subject and predicate segments, and presented with a 200 ms interval between them. In the predicate segment, they found a small difference in the N200 for true versus false for ‘some’ sentences, but not for ‘most’ sentences. Furthermore, there were no differences in the N400, and both elicited a post-N400 positivity for false versus true trials, which lasted until the end of the trial for ‘most’, but not for ‘some’. In the subject segment, a significant positivity was found for ‘most’ relative to ‘some’, visible from around 300 ms and sustained throughout the epoch.

Summing up, previous studies have shown that truth value relative to a picture does elicit the same truth value effects as verification tasks without pictorial material, i.e., larger N400s for false than for true sentences. These N400s do not arise before the truth value of the sentence can be confidently determined, and they are followed by an increased positivity when the complexity of sentence-picture matching places greater cognitive demands on the decision process. Furthermore, sustained effects are observed earlier in the sentence, indicating that verification affects the processing of the entire sentence, and not just the final disambiguating word. This is true regardless of whether the complexity stems from the picture or the sentence.

### 1.3. *The present study*

In two ERP experiments, we sought to determine whether differences in the computational complexity of the verification algorithm for different quantifier classes are reflected online during sentence processing. Notably, proportional quantifiers should be computationally more demanding, in terms of the neural responses they elicit, than nonproportional quantifiers, here Aristotelian and numerical quantifiers (Baggio, 2018; Baggio & Bremnes, 2017). The complexity differences between proportional and nonproportional quantifiers should be reflected in real-time ERP signals in an explicit verification task, and not when participants are only asked comprehension questions.

Importantly, this question is on a higher level of abstraction than the one posed in a parallel behavioral literature, investigating specific algorithms associated with specific quantifiers (Hackl, 2009; Hunter et al., 2017; Knowlton et al., 2021; Lidz et al., 2011; Pietroski et al., 2009; Pietroski et al., 2011; Talmina et al., 2017; Tomaszewicz, 2011). The formal proofs outlined above demonstrate that, regardless of which specific algorithm is implemented to verify a proportional quantifier, the algorithm still minimally requires a push-down automaton (PDA) with a memory component to perform the task, thereby making it more computationally complex than the corresponding finite state automaton (FSA) algorithms for the nonproportional quantifiers. Relatedly, the notion of memory evoked by the automata theory is also highly abstract. The implication of specific types of memory resources employed by the brain, and therefore of specific ERP components associated with them, is not strictly predicted by the theory, and as such remains an open empirical question not addressed by the experiments presented herein.

In the present study, participants saw images of red or yellow circles and triangles, and subsequently read quantified sentences about the contents of the picture. In the first experiment, participants had to judge whether the sentence was true or false of the picture, and in the second, they had to answer comprehension questions about the picture, the sentence, or both.

We expect false sentences to elicit a sentence-final N400 type of response. If that is observed, we can reasonably conclude that the sentence has been processed and understood. Furthermore, if effects of truth value are indeed detected, we can also infer that, at that stage, the verification algorithm has already been executed. Possible ERP differences resulting from algorithmic complexity must then be observed prior to the onset of the truth value effect. To establish that these effects are related to the verification procedure, we must rule out that these differences stem from other sources, in particular comprehension processes. Thus, if different ERP effects between quantifier classes are observed only in experiment 1 (verification) but not in experiment 2 (comprehension), then they can be hypothetically considered as candidate neural signatures of the algorithmic processes posited by the formal theory.

## 2. Experiment 1

### 2.1. *Method*

#### 2.1.1. *Design*

We used a 3 × 2 design with the factors Quantifier Class (3 levels: Aristotelian, Numerical, and Proportional) and Truth Value (2 levels: True and False). Participants performed a picture-sentence verification task for each trial. To prevent eye movements that would affect the EEG recording, participants could not look at the picture while the sentence was presented and verified. Instead, a picture was shown before each sentence, at the beginning of each trial. To ensure that participants could memorize the picture well enough, and that memory encoding or recall of the picture as such would not interfere with deployment of memory resources for verification, the same picture was used within a block. Additionally, participants had the opportunity to study the picture as long as they wanted at the beginning of each block. Details on stimulus presentation, block design, and task are given below.

In this experimental set-up, all quantifier classes require some form of memory in order for participants to perform the task. However, the automata theory shows that verification of proportional quantifiers further requires manipulation of items in memory, specifically comparing two sets of objects: this requires an additional memory component. This is predicted to further increase memory load, as compared to the other two classes.

#### 2.1.2. *Participants*

Thirty right-handed native Norwegian speakers (13 female; mean age 21.53, SD = 2.58; age range 18–27), with normal or corrected to normal vision and no psychiatric or neurological disorders, were recruited from the local student community. Twenty-four participants (11 female; mean age 21.65, SD = 2.73; age range 18–27) met the inclusion criteria of having an average of at least 20 artifact-free trials per condition, and were included in the final analysis. All participants gave written informed consent and were compensated with a voucher. The study was approved by *The Norwegian Centre for Research Data* (NSD; project nr. 455334).

### 2.1.3. Materials

Twelve images consisting of clusters of 2–5 red and yellow circles and triangles in a 2 × 2 grid were constructed. The colors red and yellow were chosen because their color words both end in consonants in Norwegian ('rød' and 'gul', respectively), and preference for plural '-e' congruence marking on color words ending in vowels varies within the population (Faarlund et al., 1997, p. 370). The location, number, and color of the shapes were varied pseudorandomly. Importantly, we chose to vary both shape and color to guarantee that participants could not know the truth value of the sentence before the final word. Previous experiments with similar set-ups (e.g. Brodbeck et al., 2016) have all emphasized the need for simple pictures from which quantity information can be rapidly extracted to minimize memory encoding and subsequent retrieval. This is particularly important since quantifier class is expected to modulate memory, and such effects would be hard to detect if memory load was already high in all conditions. Note that the hypothesis above, derived from the formal proofs, is that proportional quantifiers are more difficult and require a memory component regardless of the cardinality of the set of objects: there is no strategy that can simplify the task.

To construct the sentences, two quantifiers from each quantifier class were chosen. Consequently, 6 different quantifiers were used in the stimulus set. In order to maintain syntactic identity between sentences, only quantifiers that take a plural definite complement were chosen. Numerical quantifiers were 'tre av' (*three of*) and 'fem av' (*five of*), and the Aristotelian quantifiers were 'alle' (*all*) and 'ingen av' ('none of'). 'Some' was not chosen because it affords two interpretations: a logico-semantic *at least one* reading and a pragmatic *some but not all* reading (e.g. Levinson, 1983, p. 134). For proportional quantifiers, 'de fleste' (*most*) and 'færrest av' (*the fewest*) were chosen. Downward monotone quantifiers are less frequent than upward monotone (Szymanik & Thorne, 2017), but since we wanted the two quantifiers to have complementary truth values, we decided to include 'færrest av'. Another issue with the proportional quantifiers, is that 'de fleste', like 'most' (e.g. Hackl, 2009), has both a proportional and a superlative/comparative meaning, whereas 'færrest av' does not. However, since the two meanings are denotationally equivalent in binary contexts, when there are only two alternatives, this issue was ignored. It is also important to note that 'færrest av' – in contrast to its English translation – takes a definite complement, and thus behaves identically to all the other quantifiers with respect to predicating a property of a set of objects. For an overview of the semantics of quantity adjectives in Germanic languages, and in particular the differences between the Scandinavian languages and English with respect to definiteness, see Coppock (2019).

All sentences had the form of quantifier + shape noun + copula + color adjective, see Table 1. Each quantifier was presented equally many times with all shape and color combinations in a total of 288 sentences (48 per quantifier and 96 per quantifier class). The sentences were counterbalanced according to truth value between each of twelve blocks with 24 trials each. Because the image remained the same within a block, some sentences occurred more frequently in some blocks than in others, and the ratio of true to false sentences differed slightly between blocks (*range*: 9–14; *median*: 12.5), but were evenly balanced through the experiment overall. The order of the sentences were randomized within each block. Further, we created 2 randomizations of the order of the blocks, and these were run both forward and backward, resulting in 4 different orders of the blocks, to ensure that training effects were distributed equally across trials: the imbalance of sentence-types in the different blocks was counterbalanced by participants encountering them at different stages of the experiment in random order.

All pictures and sentences can be found in the supplementary material.

**Table 1**
Experiment sentences.

| Quantifier | Shape | Copula | Color |
|---|---|---|---|
| De fleste | | | |
| *Most of* | | | |
| Færrest av | sirklene | | røde |
| *The fewest of* | *the circles* | | *red* |
| Tre av | | | |
| *Three of* | | er | |
| Fem av | | *are* | |
| *Five of* | | | |
| Alle | trekantene | | gule |
| *All of* | *the triangles* | | *yellow* |
| Ingen av | | | |
| *None of* | | | |

### 2.1.4. Procedure

After reading the information sheets and signing the consent forms, participants were seated in front of a computer screen in a dimly lit, sound attenuated, and electrically shielded EEG booth. They were instructed to judge whether each sentence was true or false of the picture seen before each trial by using two predefined response buttons (Fig. 1). Which button indicated true or false was counterbalanced between blocks, and participants were informed of this by two squares with the words 'sant' (*true*) and 'usant' (*false*) on horizontally opposing sides of the screen, with the alternatives on the side of the screen corresponding to the relative placement of the response keys. This information was provided both at the beginning of the block and every time they had to make a truth value judgement. As numerical quantifier interpretation is known to vary between participants, they were asked to interpret these exactly (e.g., *three and no more than three*) rather than as a lower bound (e.g., *at least three*). It was especially important to ensure that all participants interpreted the sentences in the same way, because the two readings have been shown to give rise to different ERP profiles (Spychalska et al., 2019). The choice of the *exact* reading was made on the grounds that this reading is preferred by the majority of people (Shetreet et al., 2014; Spychalska et al., 2019). Finally, they were told not to blink or move while reading the sentences, and that any necessary such activity could take place only while looking at the picture or when they saw a fixation cross.

At the beginning of each block, after the indication of which buttons corresponded to true and false was provided, participants saw the picture that would be presented before each trial in that block. They were advised to study the picture carefully and press a button when they were ready to begin. Each trial began with the presentation of the picture for 4 s. The picture was followed by a 500 ms fixation cross and 500 ms of blank screen. Subsequently, the sentence was presented one word at a time for 400 ms with a 400 ms blank screen onset delay. The quantifier was always presented as one expression and on a single screen frame, even if it was not a single syntactic word. This was done in order to make the length of every trial identical, which was necessary to be able to compare verification procedures. After the sentence had been presented, the same fixation cross and blank screen followed, before participants had to press a button to indicate whether the sentence was true or false. Once they had responded, or if they had not responded for 4000 ms, a new trial started immediately. When they had completed all 24 trials in the block, the experiment was paused and the participant had to press a button to begin the next block. Consequently, participants were free to determine the length of the break themselves. Each experimental session lasted between 1:10 and 1:20 hours, including breaks.

### 2.1.5. EEG-recording

EEG signals were recorded from 32 active electrodes (Fp1, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, TP9, CP5, CP1, CP2, CP6, TP10, P7, P3, Pz, P4, P8, PO9, O1, Oz, O2, and PO10), using the actiCAP system by Brain Products GmbH. The implicit reference was placed on the left mastoid, and all channels were re-referenced off-line to the averaged mastoids. EEG data were sampled at 1000 Hz using a 1000 Hz high cutoff filter and a 10 s time constant. Impedance was kept below 1 kOhm across all channels throughout the experiment.

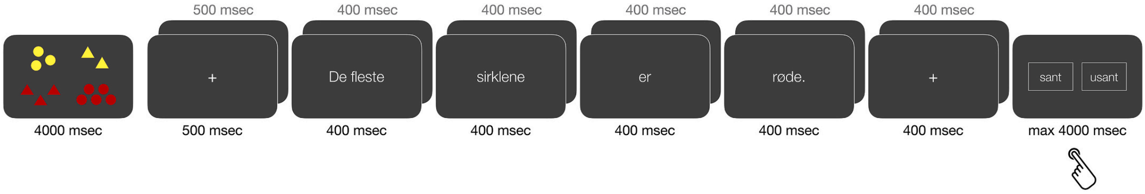


**Fig. 1.** Structure of a single trial from experiment 1. Trial structure was the same in experiment 2, except that the true/false (sann/usann) screen was replaced by a comprehension question (4000 ms) followed by a maximum 4000 ms interval within which the participant could produce an answer.

### 2.1.6. Data analysis

Accuracy and reaction time data were collected. The principal function of accuracy in this experiment was to ensure that participants were actually correctly verifying the sentences. Reaction times were primarily gathered in order to compare our study to previous behavioral experiments, but as there was a 1400 ms delay between the presentation of the final word and the response due to the fixation cross, it was acknowledged that they would not be directly comparable. The accuracy and reaction time data were subjected to mixed effects logistic and linear regression, respectively, using the glmer function of the lme4 package (Bates et al., 2015) in R. Quantifier class and truth value were fixed effects and the models had random intercepts by participant. We did not include random intercepts by item, since aside from the experimental manipulation (i.e. replacing the quantifier) the experimental stimuli were identical. As a consequence, the variance between items is not random, but is captured by a fixed effect. For both fixed effects, model comparison was performed.

EEG data were analyzed using FieldTrip (Oostenveld et al., 2011). At the quantifier, at the noun completing the noun phrase, and at the sentence-final adjective, 1000 ms epochs were extracted, including a 200 ms prestimulus interval that was used for baseline correction, and re-referenced to the averaged mastoids. Using automated artifact rejection, any trial in which one or more electrodes exceeded ±150 μV relative to baseline were rejected. Additionally, trials including eye movements were excluded by thresholding the z-transformed value of the preprocessed raw data from Fp1 and Fp2 in the 1–15 Hz range. The remaining trials were subsequently low-pass filtered at 30 Hz. Participants that had an average of fewer than 20 out of 24 trials per condition were excluded from the analysis. 6 participants did not meet these criteria.

ERPs were computed for each sentence segment by averaging all trials in one condition, that is, a sentence segment by quantifier by truth value. The same procedure was used to compute ERPs for collapsed conditions: sentence segment by quantifier class, truth value at the final word, and quantifier class by truth value at the final word. Numerical and Aristotelian quantifiers were computed both as individual classes and as a collapsed class. Because the quantifier was presented in a single frame, quantifiers differed both in length, frequency, and to a certain extent morphology and syntax: any differences here might be caused by small saccadic eye-movements, frequency, or ease of comprehension. In order to avoid these confounds, we only analyzed the parts of the sentence where participants were presented with identical linguistic material, so that the only difference between them was based on the algorithm being computed.

The ERPs were analyzed using non-parametric cluster-based statistics (Maris & Oostenveld, 2007), with alpha thresholds at 0.05 for both sample and cluster level. To assess differences between conditions, each channel-time pair (or sample) in two conditions were compared by means of a *t*-test. If the results of this test were significant at the 0.05 alpha level in at least 2 neighbouring channels and 2 neighbouring time-points, these channel-time pairs were made into a cluster, and the *t*-values of all channel-time pairs were summed. To assess statistical significance at the cluster-level, *p*-values were estimated using Monte Carlo simulations. In a cluster, all participant level channel-time pairs across conditions were collected into a single set which was then randomly partitioned into two subsets. This procedure was repeated 1000 times. The *p*-value was estimated by the number of partitions in which the test statistic was larger than in the observed data. In each case, the output is a set of (possibly empty) spatio-temporal clusters in which a pair of conditions are significantly different: we report the $T_{sum}$, size (S) and estimated *p*-values in the highest-ranked clusters. For additional details, see Maris and Oostenveld (2007).

## 2.2. Results

### 2.2.1. Behavioral results

Overall accuracy was high (mean = 0.945, SD = 0.229), and even within groups all means were above 0.9 (see Table 2 for descriptive statistics). When fitted to a mixed effects logistic regression model with accuracy as a binomial dependent variable and random intercepts by participants (see Table 3), $\beta$ estimates revealed that participants were significantly ($p < 0.0001$) less accurate with both proportional and numerical quantifiers relative to Aristotelian quantifiers. The effect of truth value was not significant ($p = 0.9$). We then re-fitted the models without one of the fixed effects, and we compared the re-fitted models to the full models by means of an ANOVA. Removing condition led to a significantly poorer model ($\chi^2 = 103.17$, $p < 0.0001$), whereas removing the effect of truth value did not significantly impact model fit.

Response times were fast both in general (mean = 659.8 ms, SD = 566.6) and across quantifier classes (see Table 2). A mixed effects linear regression model was fitted to the data with random intercepts by participants (see Table 4). It revealed a significant increase in reaction time for numerical ($p = 0.005$) and proportional ($p < 0.0001$) quantifiers relative to Aristotelian quantifiers. True sentences also elicited significantly ($p = 0.035$) faster responses than false sentences. Results of the same type of model comparison as for the logistic regression above, indicated that both quantifier class ($\chi^2 = 23.34$, $p < 0.0001$) and truth value ($\chi^2 = 5.194$, $p = 0.023$) contributed to explaining the variance in reaction time.

### 2.2.2. EEG results

*2.2.2.1. Sentence-final effects: adjective.* We first consider ERP effects at the sentence-final adjective. This is the earliest point in time at which participants can determine with confidence whether a sentence is true or false. We therefore expect that neural responses at the adjective will show sensitivity to truth value. Overall, false trials show a more

**Table 2**
Accuracy and response times, Experiment 1.

| | Accuracy | | Response time | |
|---|---|---|---|---|
| Quantifier class | Mean | SD | Mean | SD |
| Aristotelian | 0.979 | 0.143 | 623.3 | 507.7 |
| Numerical | 0.915 | 0.279 | 662.7 | 575.4 |
| Proportional | 0.939 | 0.238 | 694.0 | 610.6 |

**Table 3**
Logistic regression on accuracy, Experiment 1.

| Condition | $\beta$ | SE | $z$ | $p$ |
|---|---|---|---|---|
| Intercept | 3.9402 | 0.1819 | 21.659 | < 0.0001 |
| Numerical | −1.4870 | 0.1636 | −9.092 | < 0.0001 |
| Proportional | −1.1107 | 0.1698 | −6.540 | < 0.0001 |
| True | 0.0134 | 0.1065 | 0.126 | 0.9 |

**Table 4**
Linear regression on response times, Experiment 1.

| Condition | $\beta$ | SE | $t$ | df | $p$ |
|---|---|---|---|---|---|
| Intercept | 638.291 | 54.267 | 11.762 | 24.84 | <0.0001 |
| Numerical | 41.847 | 15.019 | 2.786 | 6806.99 | 0. 0054 |
| Proportional | 72.404 | 15.042 | 4.813 | 6806.99 | <0.0001 |
| True | −27.991 | 12.282 | −2.279 | 6806.99 | 0.0227 |

negative-going complex ERP response than true trials, largely similar across quantifier classes (Fig. 2). Statistical analyses of ERP effects in the comparison between false and true trials, collapsing across quantifier classes, show a large negative cluster between 200 and 500 ms from adjective onset with a broad scalp distribution (first-ranked cluster, NEG1: $T_{sum} = -28189.93$, $S = 5631$, $p < 0.001$) and a smaller negative cluster between 600 and 800 ms (second-ranked cluster, NEG2: $T_{sum} = -6246.91$, $S = 2123$, $p = 0.019$; Fig. 3). The effect is also present for each quantifier class taken separately (Aristotelian, first-ranked cluster, NEG1: $T_{sum} = -41153.75$, $S = 10532$, $p < 0.001$; numerical, first-ranked cluster, NEG1: $T_{sum} = -15925.43$, $S = 4123$, $p = 0.002$; proportional, first-ranked cluster, NEG1: $T_{sum} = -6389.83$, $S = 2136$, $p = 0.012$; Fig. 3). These were the only clusters in which the associated Monte Carlo $p$-values are below the $\alpha = 0.05$ threshold. The decreasing cluster sizes (S) and cluster-level $T_{sum}$ statistics from Aristotelian to numerical to proportional indicate that the size of the truth value effect in ERPs varies accordingly, with the largest effect observed for Aristotelian quantifiers and the weakest for proportional quantifiers.

An inspection of ERP waveforms (Fig. 2) provides further information on the nature of these effects and their possible underlying physiology. ERP waveforms do not differ between conditions in the first 200 ms after adjective onset, up to and including the N100-P200 complex. From about 200 ms, waveforms differ qualitatively between false and true trials, and these qualitative differences are modulated by the

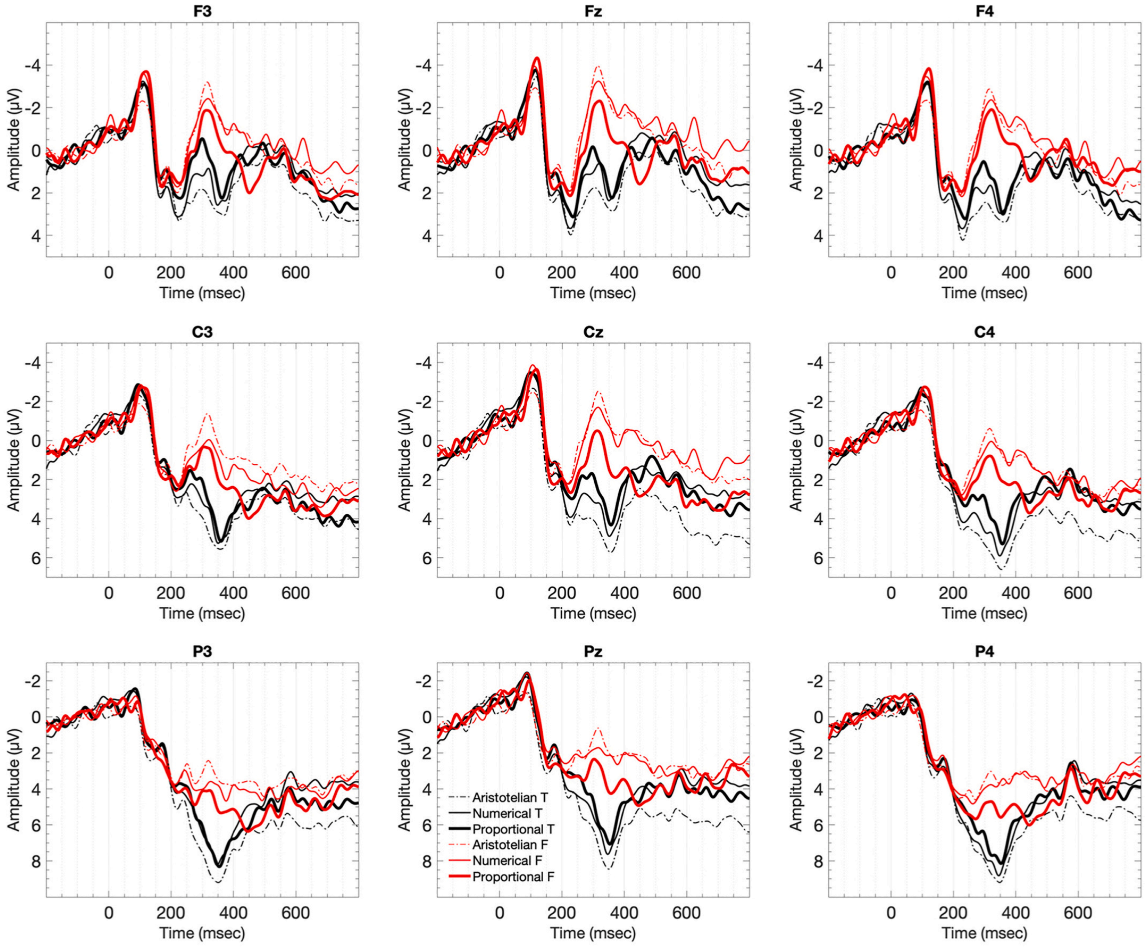


**Fig. 2.** Grand-average ERP waveforms from 9 selected channels, time locked to the onset of the sentence-final adjective (0 ms) in experiment 1. True trials are shown in black, false trials in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

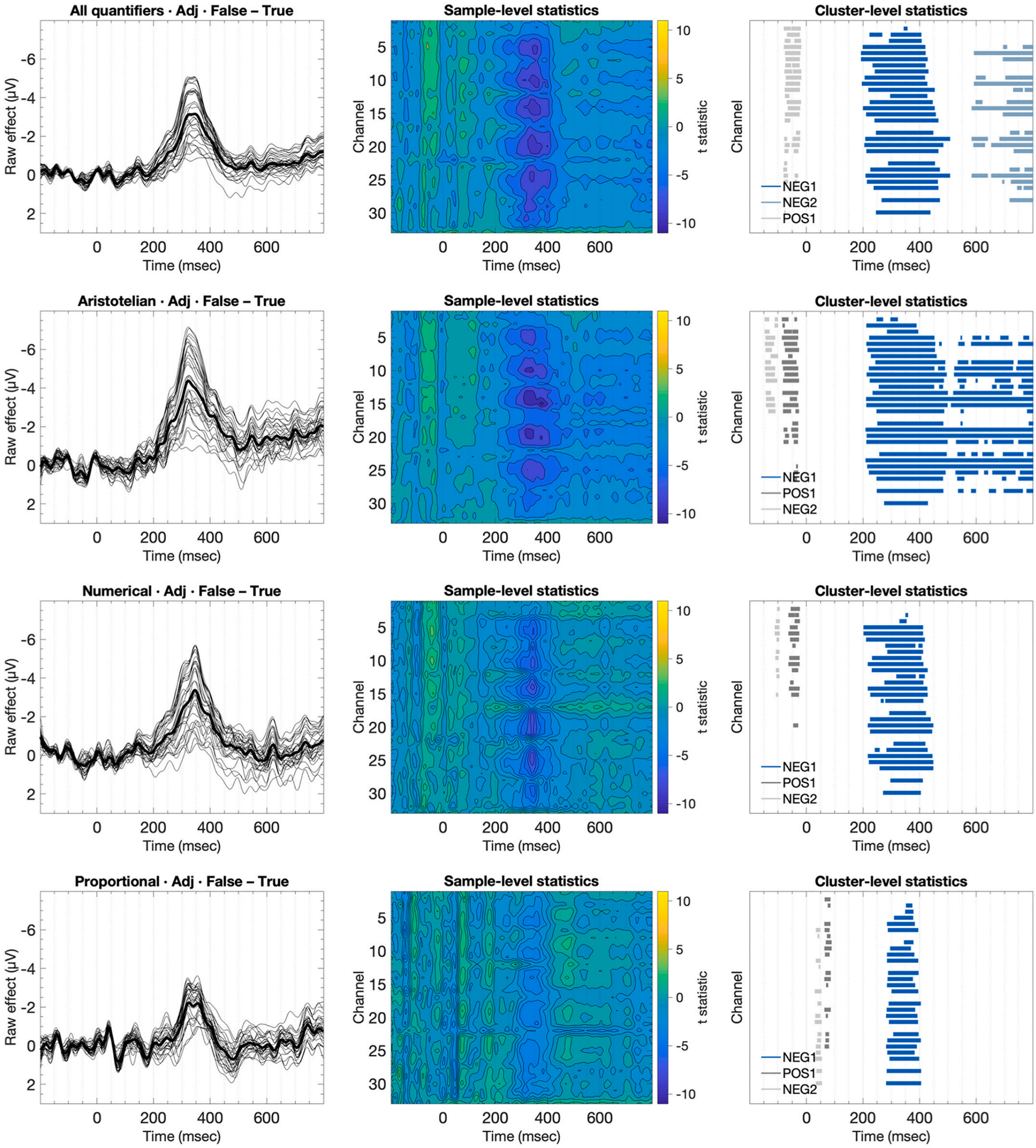


**Fig. 3.** ERP effects of truth value (False-True) across quantifier classes, time locked to the onset of the sentence-final adjective (0 ms) in experiment 1. Raw effect waveforms (left column) are displayed along with contour maps of sample-level statistics (middle column) and raster plots of cluster-level statistics (right column). Clusters with an associated $p$-value below the specified threshold ($\alpha = 0.05$) are shown in blue shades; all other clusters (gray shades) were statistically not significant. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

quantifier classes. All true trials present a clear P300 component, particularly visible over posterior channels (Fig. 2, black lines). The P300 component appears largest for true trials with Aristotelian quantifiers and smallest for true trials with proportional quantifiers, with numerical quantifiers falling in between. These differences persist throughout the epoch (Fig. 2). In direct comparisons between true trials across quantifier classes, we only found a marginal effect for the first-ranked cluster in the contrast between Aristotelian and proportional quantifiers ($T_{sum} = 2081.42$, $S = 806$, $p = 0.072$), and no effects for Aristotelian vs numerical or numerical vs proportional. These data indicate that verification strategies at the sentence-final word for true trials do not differ, in terms of underlying physiology, between quantifier classes.

ERP waveforms appear qualitatively different in false trials. All false trials present a visible rising flank of the N400 component (Fig. 2, red lines) or possibly of an N200-N400 complex. After 300 ms from adjective onset, waveforms from false trials show a positive-going deflection: this coincides temporally with the P300 in true trials, suggesting that a P300 wave may overlap with the peak and the falling flank of the N400 component, rendering its characteristic features less visible here. Importantly, from around 300 ms, the waveforms for false trials diverge between the quantifier classes. They pattern together in false trials with Aristotelian and numerical quantifiers, showing more negative voltage values overall and no differences between them (no positive or negative clusters with a significant effect). Differences were found between Aristotelian and proportional quantifiers (first-ranked cluster: $T_{sum} = -5013.65$, $S = 1635$, $p = 0.015$) and between numerical and proportional quantifiers (first-ranked cluster: $T_{sum} = -3969.17$, $S = 1394$, $p = 0.034$), indicating that proportional quantifiers are associated with a more positive-going deflection in ERPs than both Aristotelian and numerical. These results suggest that verification strategies at the sentence-final word for false trials differ, in terms of underlying physiology, between proportional quantifiers and Aristotelian-numerical quantifiers.

*2.2.2.2. Sentence-internal effects: noun.* We now consider ERP effects at the sentence-internal noun position. This is the earliest point in time at which participants can effectively initiate the verification process, recalling from memory the content of the picture, storing in memory the content of the sentence, and integrating the two. We therefore expect that neural responses at the noun will show sensitivity to the computational complexity of the different quantifier classes, with proportional quantifiers resulting in qualitatively different ERP responses than Aristotelian and numerical quantifiers. At the noun, we observed diverging ERP responses between the quantifier classes following the N100-P200 complex. Numerical quantifiers exhibit a more negative-going ERP response throughout the epoch, proportional quantifiers elicit a more positive-going response, and Aristotelian quantifiers tend to fall between the two (Fig. 4). Direct comparisons between numerical and Aristotelian quantifiers reveal only a marginal ERP effect in one small negative

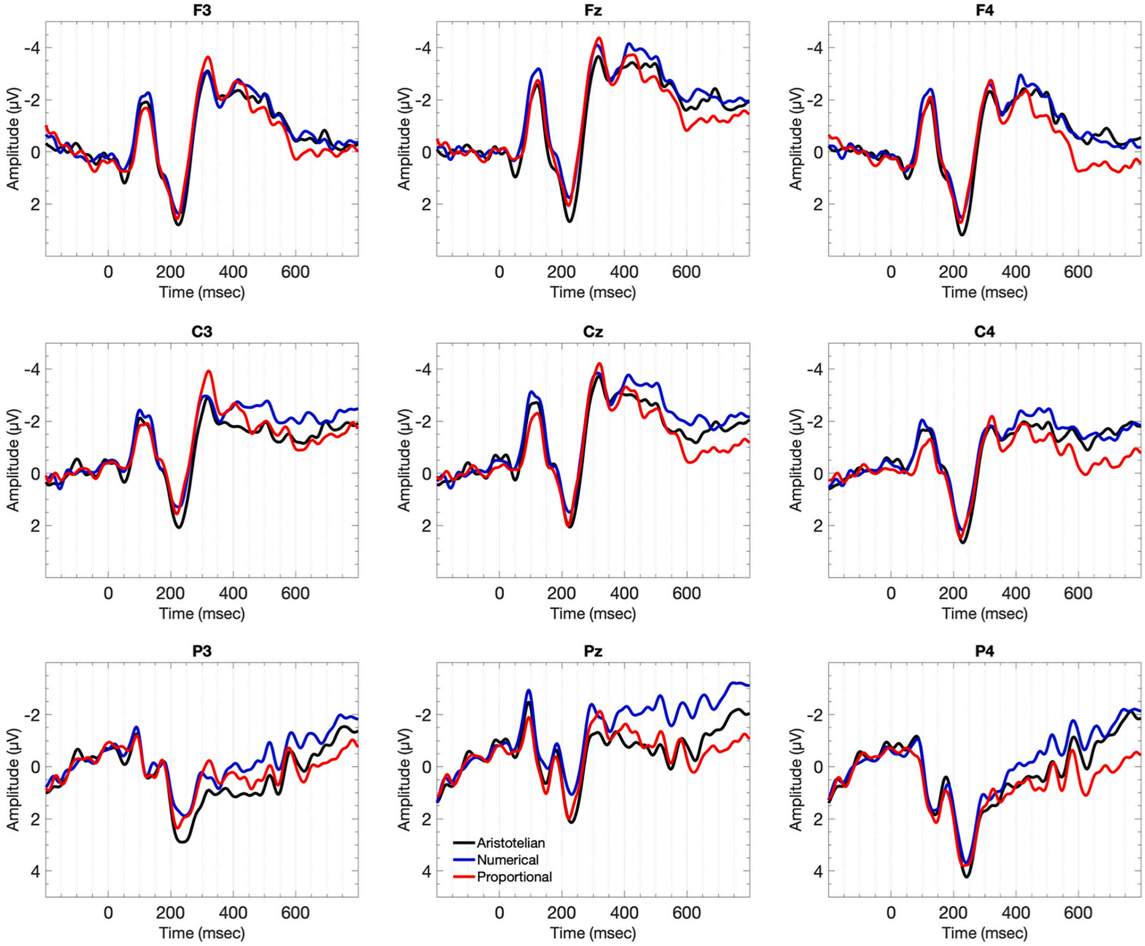


**Fig. 4.** Grand-average ERP waveforms from 9 selected channels, time locked to the onset of the sentence-internal noun (0 ms) in experiment 1. Trials from nouns following Aristotelian quantifiers are shown in black, blue is numerical quantifiers, and red is proportional quantifiers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cluster (first-ranked cluster, NEG1: $T_{sum} = -2193.62$, $S = 814$, $p = 0.081$; Fig. 5). In contrast, we found larger positive clusters in the comparisons between proportional and Aristotelian quantifiers (first-ranked cluster, POS1: $T_{sum} = 3183.25$, $S = 1237$, $p = 0.041$), proportional vs numerical quantifiers (first-ranked cluster, POS1: $T_{sum} = 3231.82$, $S = 1177$, $p = 0.040$), and proportional vs numerical and Aristotelian collapsed (first-ranked cluster, POS1: $T_{sum} = 5888.53$, $S = 2225$, $p = 0.019$; Fig. 5). This positive ERP shift, driven by proportional quantifiers relative to the two other classes, is largest after 600 ms from noun onset, both in terms of voltage values and statistically. Its

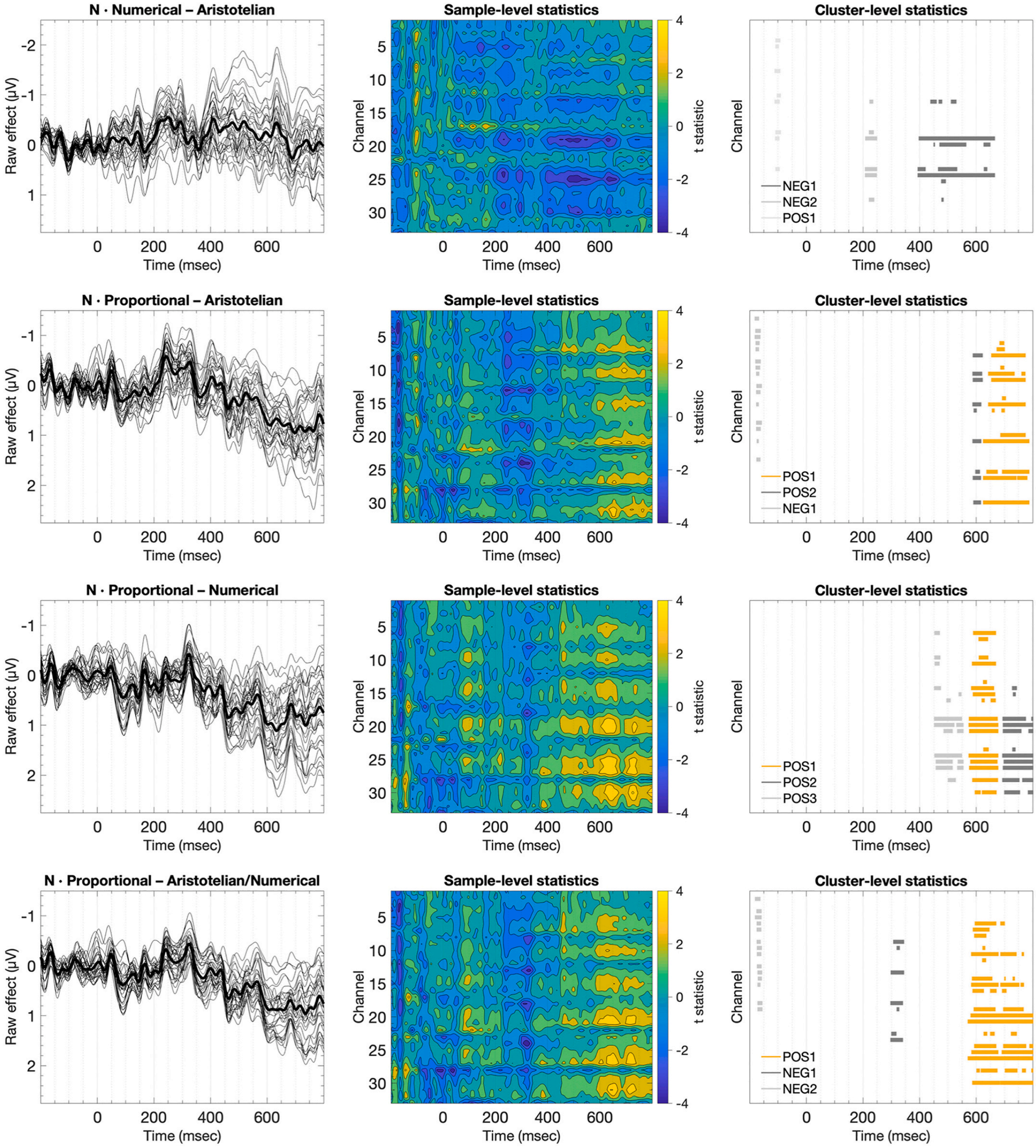


**Fig. 5.** ERP effects of pairwise comparisons between quantifier classes, time locked to the onset of the sentence-internal noun (0 ms) in experiment 1. Raw effect waveforms (left column) are displayed along with contour maps of sample-level statistics (middle column) and raster plots of cluster-level statistics (right column). Clusters with an associated $p$-value below the specified threshold ($\alpha = 0.05$) are shown in yellow shades; all other clusters (gray shades) were statistically not significant. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

temporal profile and posterior distribution (Fig. 5, contour plots of sample-level statistics) appear more consistent with a P600 effect than with earlier positivities, such as the P300.

### 2.3. Interim discussion

The sentence-final negative effect of truth value revealed that participants are correctly performing the task. The negativity was also modulated by Quantifier Class, such that the largest effect was found for Aristotelian and the smallest for proportional, with numerical quantifiers in between. Furthermore, while there were no significant differences between the classes in true trials, proportional quantifiers differed from the other two in false trials. Notably, we observed that, from around 300 ms, proportional quantifiers are more positive than Aristotelian and numerical. These results are comparable to the effects from Augurzky et al. (2017) in that the negative effect is somewhat earlier than a standard N400, and the condition that is predicted to be more complex gives rise to a post-N400 positivity. Since a truth value effect presupposes that a verification procedure has been performed, we have no reason to believe that these effects reflect the verification procedure while it is taking place. Rather, they are more likely an effect of verification complexity on subsequent cognitive processes, such as task-relevant attentional or decision processes (Augurzky et al., 2017; Sassenhagen et al., 2014).

If participants have already established sentence truth value at the final word, as our evidence indicates, then algorithmic verification differences should be observed earlier in the sentence. Indeed, we found that proportional quantifiers differed significantly from the other two classes, showing a broadly distributed positivity. The effect was largest for proportional quantifiers relative to the other two classes collapsed, but is also clearly observed between proportional quantifiers and Aristotelian and numerical individually. This effect appears consistent with a P600, both spatially and temporally. Because the ERP is recorded from the onset of the noun, where the participants were presented with identical linguistic material, the effect cannot stem from the noun itself. This leaves three options: it can be (1), an attentional or decision effect of the same kind observed at the final word; (2) an effect of the syntactosemantic combinatory procedure, such as building a compositional representation of the noun phrase or the sentence as a whole (Fritz & Baggio, 2020, 2021); or (3) an effect reflecting algorithmic verification differences between proportional and nonproportional quantifiers. It seems unlikely that participants would initiate decision making processes this early in the sentence – recall that such effects have previously only been observed when truth value can be unambiguously determined, and this only happens at the final word in the current set-up. Regarding (2), it has been claimed (Hackl, 2009) that ‘most’ is syntactically derived from its root adjective form ‘many’ and superlative morphology, thus creating a more complex noun phrase than the other classes, which both contain proper determiners rather than derived adjectives. If this is the case, then this could be a P600 integration or composition effect (Baggio, 2021; Brouwer & Hoeks, 2013). However, it is also consistent in distribution with the LPC, a centro-parietal positivity that peaks around 600 ms, associated with decision-relevant memory retrieval (Hubbard et al., 2019; Ratcliff et al., 2016; Rugg et al., 1998; Yang et al., 2019). This would be in line with the predictions of the automata theory, where the difference between the proportional and nonproportional quantifiers is precisely a memory process.

Despite these arguments, it is not possible to assess which of the above interpretations is the correct one just on the basis of data from experiment 1. We therefore conducted a second experiment, without an explicit verification task, to determine whether the effects persist when verification is no longer required, but participants still have to view the images and read the sentences. Importantly, if the positivity on the noun is a syntactosemantic combinatory effect, it should still be seen when participants read and comprehend the sentences. Similarly, the post-N400 decision effect on false sentence completions with proportional quantifiers should also disappear, as the complexity of the task remains constant between all three quantifier classes, and so no additional attentional demands are placed on participants.

## 3. Experiment 2

### 3.1. Method

#### 3.1.1. Participants

Twenty-seven (14 female; mean age 23.53, SD = 3.55; age range 19–34) participants were recruited from the same student community as in experiment 1. Twenty-four participants (12 female; mean age 23.21, SD = 3.46; age range 19–34) met the inclusion criteria and were included in the final analysis. All participants gave written informed consent and were compensated with a voucher. The study was approved by *The Norwegian Centre for Research Data* (NSD; project nr. 455334).

#### 3.1.2. Materials

The picture and sentence stimuli were identical to those in experiment 1, as was the order of presentation both within and across blocks. In addition, we constructed comprehension questions that concerned either the picture, the sentence or both. To ensure that participants were paying as much attention to both types of stimulus, half the questions included questions about both the sentence and the picture, and the other half contained an even number of questions about either. The sentence questions were of the form ‘Er setninga en påstand om *(quantifier/adjective) shape*?’ (*Is the sentence a claim about (quantifier/adjective) shape?*), whereas the questions about the picture asked ‘Er det *adjective shape* på bildet?’ (*Are there adjective shape in the picture?*). The questions about both were of the same form as the picture questions, but with the possible omission of the adjective: ‘Er det *(adjective) shape* både på bildet og i setninga?’ (*Are there (adjective) shape both in the picture and in the sentence*). Importantly, the questions about the picture and about both the picture and the sentence could not contain reference to the quantifier, as this could trigger explicit verification of the sentences. This meant that there was more variation in the questions about the sentence, than in the other two categories. The questions were balanced according to truth value and distributed evenly across the quantifier classes. However, like in experiment 1, due to the nature of the images, it was not possible to balance the truth value within each block completely, nor avoid repeating the same questions multiple times for some images. All questions can be found in the supplementary material.

#### 3.1.3. Procedure

The procedure replicated as much as possible the procedure in experiment 1. Participants sat in the same booth and used the same response buttons, received the same information at the beginning of each block, and had the same opportunity to take breaks. They also received the same instructions prior to the experiment, but the explanation of the task necessarily differed. The block and trial structure was essentially the same except that, after the sentence was presented, participants saw the comprehension question for 4000 ms, before they had to answer it with the same time-constraint as in experiment 1. This meant that the experimental sessions took approximately 20 min longer.

#### 3.1.4. EEG-recording

There were no differences in EEG recording between experiments.

#### 3.1.5. Data analysis

EEG data were processed and analyzed in the same fashion as in experiment 1. For the behavioral data, we constructed comparable mixed effects logistic and linear regression models as in experiment 1, for the accuracy and reaction time data, respectively. The only difference was that, in addition to quantifier class and sentence truth value, the question type – about the picture, the sentence, or both – and whether the question required an affirmative or negative answer, were added as fixed effects.

### 3.2. Results

#### 3.2.1. Behavioral results

Also in this experiment accuracy was high (mean = 0.934, SD = 0.247). A mixed effects logistic regression model with accuracy as a binomial dependent variable, random intercepts by participant and question type, question truth value, quantifier class and sentence truth value as fixed effects were fitted to the data. The model revealed that participants were significantly ($p < 0.0001$) more accurate with questions that only concerned the picture, relative to questions about both picture and sentence, and that they were marginally more accurate ($p = 0.038$) when the sentence contained a numerical compared to an Aristotelian quantifier. All other $\beta$-estimates were not significant.

Participants also responded quickly to the comprehension questions (mean = 654.9 ms, SD = 569.8). We fitted a mixed effects linear regression with the same parameters as in the logistic regression above to the data. Reaction times were lower when the question only concerned the picture ($p < 0.0001$) or the sentence ($p = 0.003$) compared to both, when the question required an affirmative as opposed to a negative answer ($p = 0.036$), and when the sentence contained a proportional rather than an Aristotelian quantifier ($p < 0.001$) (see Tables 5–7).

#### 3.2.2. EEG results

*3.2.2.1. Sentence-final effects: adjective.* In experiment 2 there is no explicit verification task. Participants had to answer questions about the picture or the sentence, and establishing the truth value of the latter was never required to perform the task. However, participants might still covertly track the truth and falsehood of sentences, to the extent that cognitive resources, not expended in the main comprehension task, are available for implicit verification. If covert truth tracking indeed occurs, ERP signals at the sentence-final adjective should still show sensitivity to truth value. Overall, collapsing over the quantifier classes, false trials result in more negative-going ERPs at the adjective than true trials. This negative cluster shows a similar temporal and spatial distribution to its counterpart in experiment 1, but is weaker statistically (first-ranked cluster, NEG1: $T_{sum} = -5204.02$, $S = 1860$, $p = 0.011$; Figs. 5 and 6). Moreover, and most importantly, it is only observed in the comparisons between false and true trials in Aristotelian (first-ranked cluster, NEG1: $T_{sum} = -2948.82$, $S = 1119$, $p = 0.040$) and numerical quantifiers (first-ranked cluster, NEG1: $T_{sum} = -3741.65$, $S = 1340$, $p = 0.018$), but not in proportional quantifiers, where the effect is absent (the three highest-ranked clusters are all positive clusters, but none has an associated $p$-value below threshold; Fig. 7). The negativity observed in experiment 1 in the contrast between false and true trials with proportional quantifiers is here not elicited. These results indicate that implicit verification, or covert tracking of the truth and falsehood of sentences, may still occur in either true or false trials, or both, with Aristotelian and numerical quantifiers, but it does not occur for proportional quantifiers.

*3.2.2.2. Sentence-internal effects: Noun.* ERP results from the sentence-final word in experiment 2 suggest that, in a comprehension task that does not require verification, participants do not compute the truth values of sentences containing proportional quantifiers. If this is correct, and if the positivity observed at the sentence-internal noun position for proportional quantifiers in experiment 1 reflects the complexity of the verification process, then that effect should disappear in the same contrast in experiment 2. That was indeed what we found at the noun position. As in experiment 1, ERP waveforms appear more negative for numerical than for Aristotelian quantifiers (Fig. 8), however there were no significant negative or positive clusters for that comparison specifically (Fig. 9). Contrary to experiment 1, where proportional quantifiers resulted in positive effects compared to both Aristotelian and numerical quantifiers, such effects are absent in experiment 2: there are no visible waveform differences between proportional quantifiers and the other two classes (Fig. 8) and no negative or positive clusters with associated $p$-values below the specified threshold (Fig. 9). These results indicate that implicit verification of sentences containing proportional quantifiers does not happen in experiment 2 (missing sentence-final effect of truth value) and is not even attempted (missing sentence-internal effect of quantifier class). These conclusions support the hypothesis that the positivities observed at the noun and at the adjective in experiment 1 reflect the computational complexity of the verification process for sentences containing proportional quantifiers.

**Table 5**
Accuracy and response times, Experiment 2.

| | Accuracy | | Response time | |
|---|---|---|---|---|
| Question type | Mean | SD | Mean | SD |
| Both | 0.920 | 0.272 | 682.8 | 606.7 |
| Picture | 0.974 | 0.158 | 614.3 | 498.4 |
| Sentence | 0.924 | 0.265 | 640.0 | 558.2 |
| Quantifier class | | | | |
| Aristotelian | 0.928 | 0.259 | 674.1 | 602.8 |
| Numerical | 0.944 | 0.230 | 666.9 | 581.2 |
| Proportional | 0.932 | 0.253 | 623.8 | 521.4 |

**Table 6**
Logistic regression on accuracy, Experiment 2.

| Condition | $\beta$ | SE | $z$ | $p$ |
|---|---|---|---|---|
| Intercept | 2.4918 | 0.1742 | 14.305 | < 0.0001 |
| Picture Question | 1.2251 | 0.1656 | 7.399 | < 0.0001 |
| Sentence Question | 0.0836 | 0.1125 | 0.743 | 0.4573 |
| Question True | 0.0693 | 0.1000 | 0.693 | 0.4886 |
| Numerical | 0.2586 | 0.1244 | 2.079 | 0. 0376 |
| Proportional | 0.0941 | 0.1187 | 0.793 | 0. 4280 |
| Sentence True | −0.0771 | 0.0996 | −0.775 | 0.4386 |

**Table 7**
Linear regression on response times, Experiment 2.

| Condition | $\beta$ | SE | $t$ | df | $p$ |
|---|---|---|---|---|---|
| Intercept | 719.479 | 46.924 | 15.333 | 27.776 | < 0.0001 |
| Picture Question | −72.638 | 15.668 | −4.636 | 6807.004 | < 0.0001 |
| Sentence Question | −46.773 | 15.711 | −2.977 | 6807.007 | 0.0029 |
| Question True | −26.989 | 12.850 | −2.100 | 6807.008 | 0. 0357 |
| Numerical | −7.147 | 15.792 | −0.453 | 6807.005 | 0. 6509 |
| Proportional | −53.745 | 15.757 | −3.411 | 6807.005 | 0. 0007 |
| Sentence True | 0.060 | 12.794 | 0.005 | 6807.014 | 0.9963 |

### 3.3. Interim discussion

We observed sentence-final negative effects for false versus true completions for Aristotelian and numerical quantifiers, albeit smaller and statistically less robust than in experiment 1. By contrast, the negativity on proportional quantifiers disappeared completely. The data therefore suggest that with Aristotelian and numerical quantifiers, participants are still able to track truth value even when not explicitly verifying the sentence, but they are not with proportional quantifiers. This may be explained by the algorithm for proportional quantifier verification being too complex to deploy when it is not strictly task relevant: the working memory resources required by the proportional verification algorithm are not available because they are allocated in the main task. This is further evidenced by the absence of sentence internal effects at the noun. An interesting side effect of participants not verifying sentences with proportional quantifiers is that it makes them faster at responding to the comprehension question. Since the more complex verification procedure is not performed at all, participants have more cognitive resources to devote to the experimental task when reading

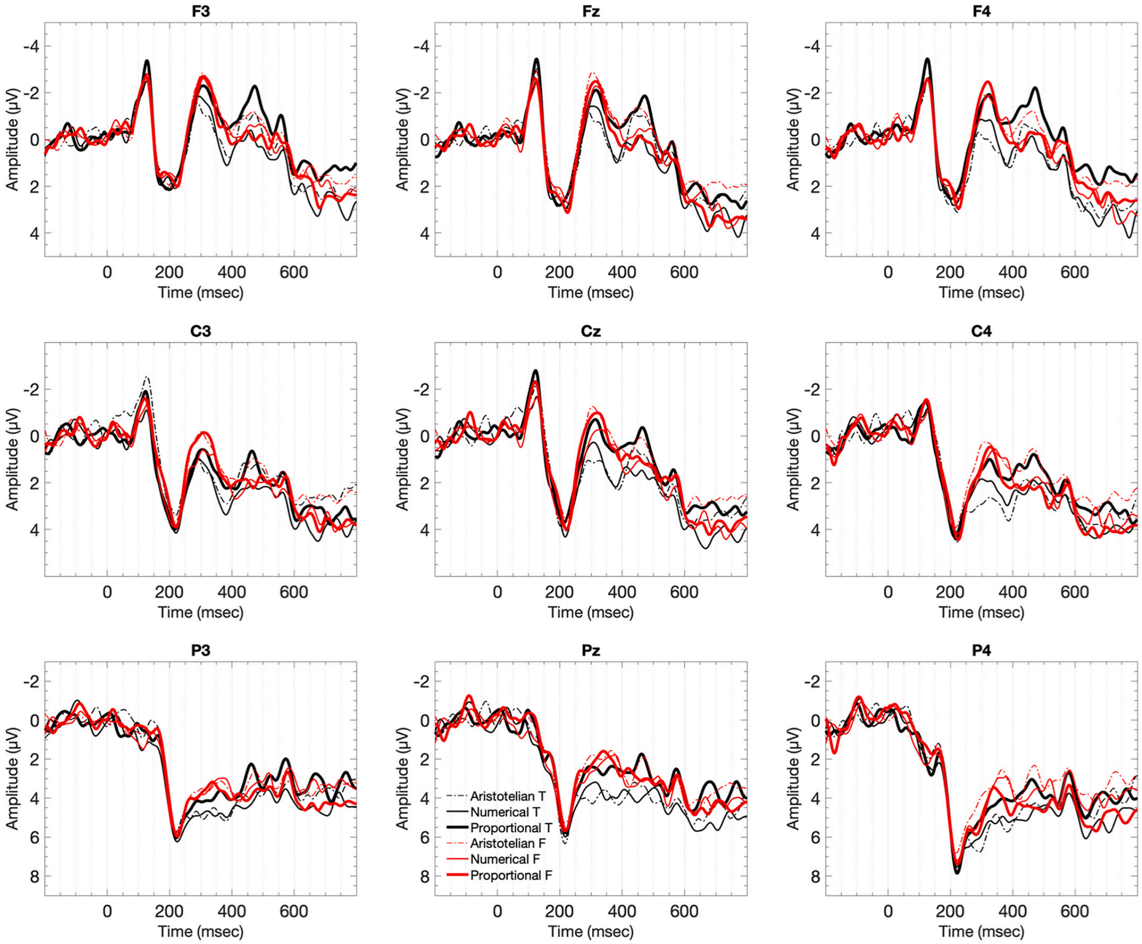


**Fig. 6.** Grand-average ERP waveforms from 9 selected channels, time locked to the onset of the sentence-final adjective (0 ms) in experiment 2. True trials are shown in black, false trials in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

proportional quantifier sentences than they do when they are simultaneously reading and verifying nonproportional sentences. This post hoc explanation of the decrease in reaction time also supports our interpretation of the cognitive process manifested in the evoked potentials. Finally, as predicted, the post-N400 positivity for proportional quantifiers in false trials also disappeared, further strengthening the view that this positivity is an attentional or decision effect.

## 4. General discussion

Overall, we found that computational complexity, as measured by algorithmic verification differences, impacts neural activity during sentence processing. When participants had to perform an explicit picture-sentence verification task (experiment 1), we found a negativity in the N200-N400 time-window at the final word. The effect of false versus true trials is larger for Aristotelian (e.g. ‘all’) than for proportional quantifiers (e.g. ‘most’), while numerical quantifiers (e.g. ‘three of’) fall in between: this finding is beyond the predictive scope of the automata theory of quantifier verification, but it shows that different quantifier classes have specific processing consequences at various stages of verification. With a comprehension question task (experiment 2), the truth value effect is attenuated for Aristotelian and numerical quantifiers, and disappear completely for proportional quantifiers. Additionally, proportional quantifiers were significantly more positive than the other two classes, both individually and collapsed, on the noun completing the subject noun phrase in the verification experiment. No such effect was found in the comprehension experiment, indicating the effect is due to verification and not to syntactosemantic differences relating to composition as per Hackl (2009).

These ERP effects can be interpreted in light of the previous literature. Most saliently, this is the same pattern observed with the auditory stimuli over pictorial contexts by De Santo et al. (2019). They found a positivity for ‘most’ relative to ‘some’ on the subject segment, and a larger positivity in false trials on the predicate segment. Importantly, we also observed differences in the size of the N200-N400 negativity, which De Santo et al. (2019) did not. This could be a power-issue, as their study only had a small number of participants, but could also be due to the mode of presentation: their participants could verify the sentence while looking at the picture, whereas our participants had to recall the image from memory. Additionally, serial visual presentation of sentences is known to elicit different neural responses than auditory stimuli (Freunberger & Nieuwland, 2016). Since no other studies have

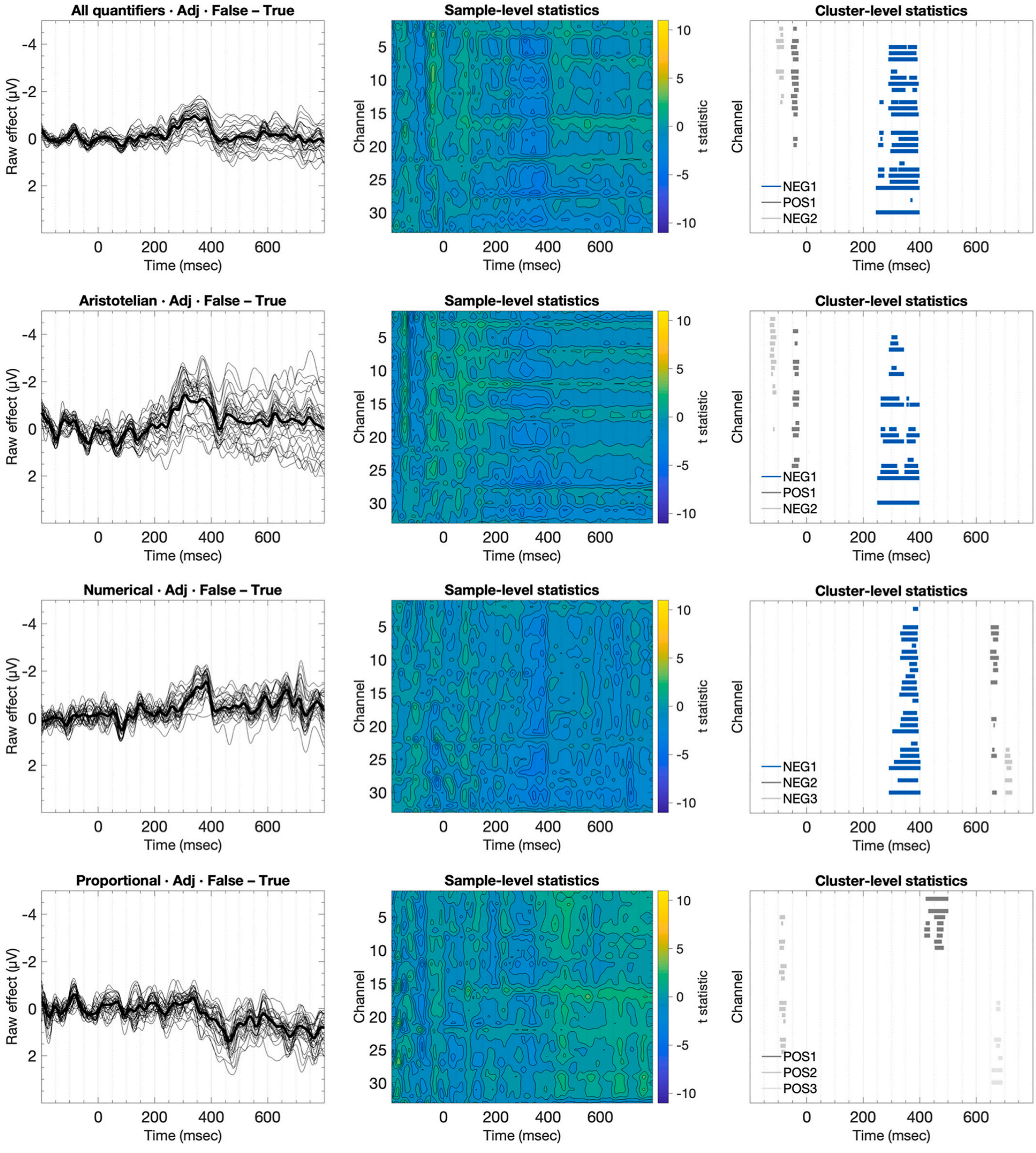


**Fig. 7.** ERP effects of truth value (False-True) across quantifier classes, time locked to the onset of the sentence-final adjective (0 ms) in experiment 2. Raw effect waveforms (left column) are displayed along with contour maps of sample-level statistics (middle column) and raster plots of cluster-level statistics (right column). Clusters with an associated *p*-value below the specified threshold ($\alpha = 0.05$) are shown in blue shades; all other clusters (gray shades) were statistically not significant. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

compared different classes of quantifiers using EEG, a graded N400 effect could not have been observed. Particularly worthy of consideration is the fact that negative quantifiers – like 'the fewest' in this study – have been found not to give rise to N400 effects (Augurzky et al., 2020a; Nieuwland, 2016; Urbach et al., 2015; Urbach & Kutas, 2010). One possibility is therefore that this is what is driving the reduced N200-N400 effect for proportional quantifiers, as this class contained both a positive and a negative quantifier. However, even if this is the case, the fact that the N200-N400 effect is graded, i.e., largest for Aristotelian, smaller for numerical, and smaller yet for proportional,

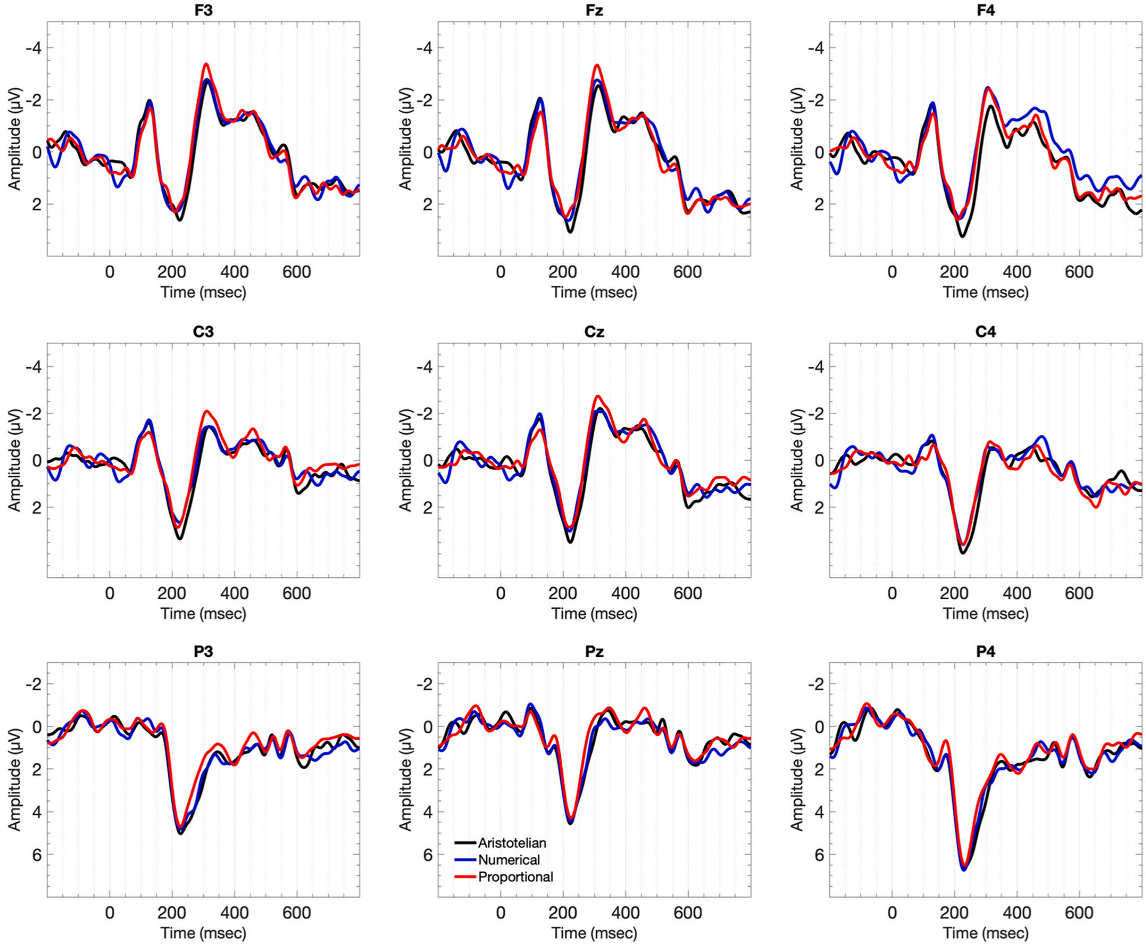


**Fig. 8.** Grand-average ERP waveforms from 9 selected channels, time locked to the onset of the sentence-internal noun (0 ms) in experiment 2. Trials from nouns following Aristotelian quantifiers are shown in black, blue is numerical quantifiers, and red is proportional quantifiers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

remains to be explained.

Another issue with the observed N200-N400 negativity is its latency. Like Augurzky et al. (2017) (see also Knoeferle et al., 2011; Vissers et al., 2008), the negativity observed for false trials is earlier than traditional N400s. It is therefore possible that it is a N2b (D'Arcy et al., 2000; Wassenaar & Hagoort, 2007), reflecting a mismatch between the active representation of the picture and the sentence. Early onset N400 effects have been demonstrated when semantic expectancy is very high (Van Petten et al., 1999), such as in the context of a picture (Vissers et al., 2008). Since both of these interpretations require the construction of a model or mental representation of the picture and the sentence, the argument made in the following does not rely on which of these interpretations turns out to be correct.

More generally, our results are consistent with and similar to previously observed ERP effect patterns. As in Augurzky et al. (2017, 2019, 2020a, 2020b), the more complex task – in our work, verifying proportional quantifiers; in their work, more complex pictorial stimuli – gave rise to a late positivity at the disambiguating position that only occurred in the verification task and that is thus plausibly related to an increase in decision complexity. The positivity at the noun also has antecedents in the literature, whether it be for semantic violations (Politzer-Ahles et al., 2013) or the increase in complexity due to negative polarity (Augurzky et al., 2020a).

Our results are best explained by a procedure in which participants are building a model verifying the sentence on-line (Baggio, 2018; Clark, 1976; Clark & Chase, 1972, 1974; Johnson-Laird, 1983; Just, 1974; Just & Carpenter, 1971; van Lambalgen & Hamm, 2005; Zwaan & Radvansky, 1998). Note that alternative explanations, for example in terms of visual context effects (Knoeferle et al., 2011; Vissers et al., 2008), also presuppose the construction of a model. This is evidenced by the N400-like negativity in false sentences relative to true, which presupposes that a verification procedure – building a model of the sentence – has taken place. Interestingly, this negativity appears to be modulated by the complexity of the verification algorithm in that the more complex the verification procedure, the smaller the negativity. As the N400 is known to be modulated by probability in a context, this could imply that participants are less able to predict, or less confident of, the final word for proportional quantifiers, an option further substantiated by the positivity following the N400 in false trials for proportional quantifiers. Crucially, this positivity can be argued to be a decision effect reflecting

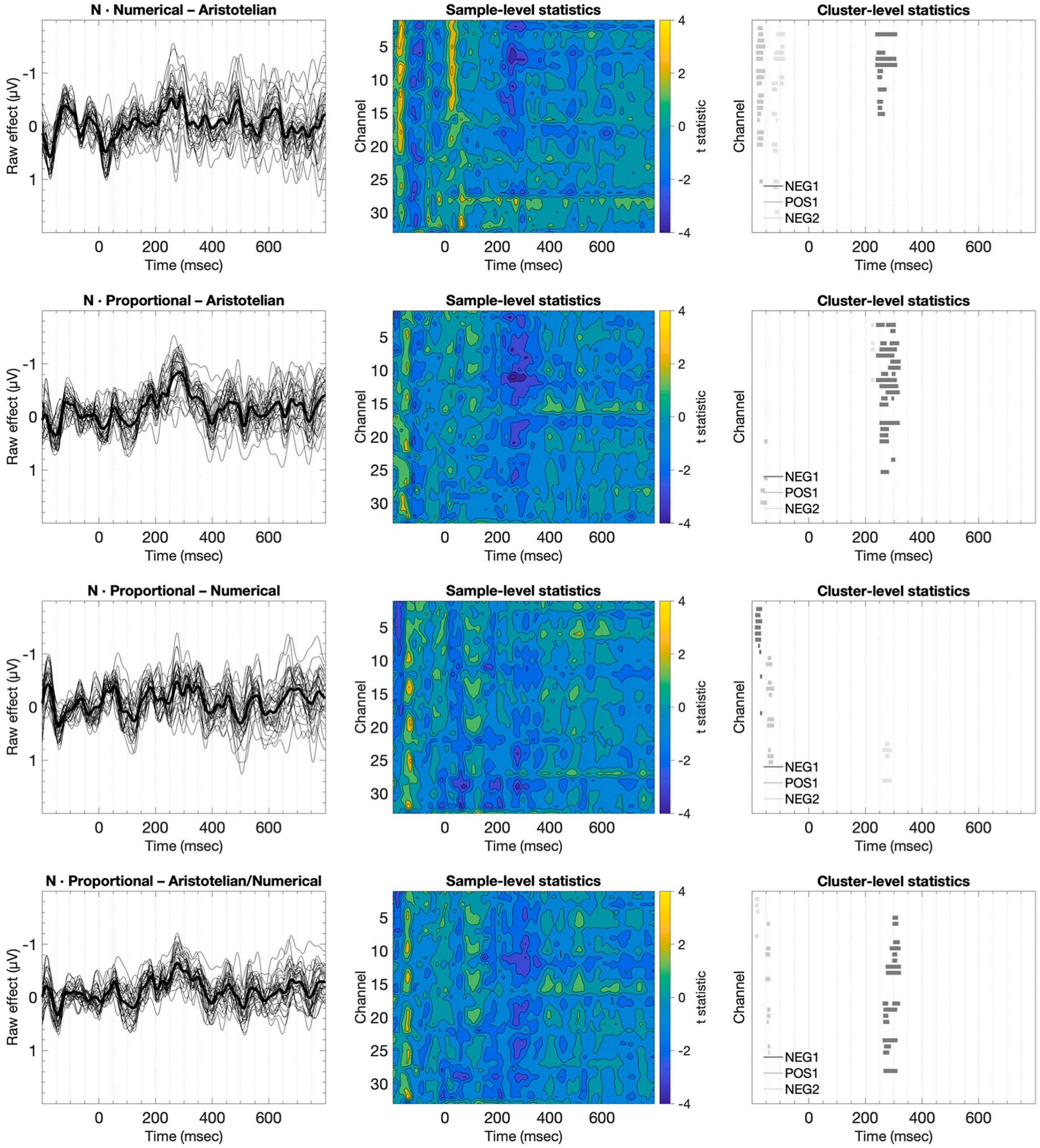


**Fig. 9.** ERP effects of pairwise comparisons between quantifier classes, time locked to the onset of the sentence-internal noun (0 ms) in experiment 2. Raw effect waveforms (left column) are displayed along with contour maps of sample-level statistics (middle column) and raster plots of cluster-level statistics (right column). Clusters with an associated *p*-value below the specified threshold ($\alpha = 0.05$) are shown in yellow shades; all other clusters (gray shades) were statistically not significant. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

increased cognitive demands (Augurzky et al., 2017; Sassenhagen et al., 2014), particularly as this effect disappears when the decision complexity is kept constant in the comprehension question experiment. The decreased certainty for proportional quantifiers may stem from the fact that more cognitive resources are required to perform the verification algorithm for proportional quantifiers, and consequently fewer resources are available for prediction.

If a model of sentence meaning has been built at the final word, then the positivity at the noun can be argued to be a signature of verification. The time-course and distribution of the effect is similar to the LPC component – often called the parietal old/new effect – from the recognition memory literature (Hubbard et al., 2019; Ratcliff et al., 2016; Rugg et al., 1998; Yang et al., 2019). The LPC is associated with recollection memory (Rugg & Curran, 2007) – i.e., when recollecting contextual details of a stimulus – and is only observed when it is task-relevant (Yang et al., 2019). Since the algorithms for proportional and nonproportional quantifiers differ precisely in the use of a memory component, an explanation in which participants recruit additional memory to perform proportional quantifier verification is well grounded in formal theory. The fact that this effect disappears along with the N400 for proportional quantifiers in the comprehension experiment further supports this interpretation. Given that a syntactosemantic composition effect would presumably manifest itself regardless of task, this explanation of the positivity at the noun is weakened by experiment 2. However, while links between P600 effects and episodic memory have been proposed (O'Rourke & Van Petten, 2011; Van Petten & Luka, 2012), this hypothesis has not been tested in actual sentence processing paradigms, but only with single words. This interpretation is therefore problematic, and there is a possibility that the positivity here indexes generic processing costs. De Santo et al.'s (2019) preliminary results, observing a similar effect when participants are listening to a sentence while viewing the picture, could be taken to support such a criticism. At the same time, the automata theory proves that, if participants go through the objects sequentially, memory resources are necessarily recruited for proportional quantifiers, but not for nonproportional quantifiers, and as such no strong conclusions can be drawn on the basis of an objection along these lines.

Regardless of the final interpretation of the observed effects, the present study demonstrates that the complexity of the verification algorithm impacts sentence processing online. Importantly, when verification is required by the task, proportional quantifiers modulate the evoked potential both when participants are constructing a true model of the sentence, as indicated by the positivity on the noun, and when this model is evaluated in relation to falsified predictions, as evidenced by sentence-final effects. On the other hand, when verification is not task-relevant, the construction of a true model that generates predictions for the final word does not occur for proportional quantifiers even though it does for both nonproportional classes.

There are some limitations of the current study. Most notably, and as mentioned above, both a sentence internal positivity and the lack of N400 effects have been observed in relation to negative polarity quantifiers (Augurzky et al., 2020a; Nieuwland, 2016; Urbach et al., 2015; Urbach & Kutas, 2010). As the current experiment did not control for polarity, it is not possible to distinguish which effects are due to negative polarity and which are due to quantifier class. To circumvent these limitations, one could firstly refer to the evidence that suggests that quantifier class also gives rise to this positive effect (De Santo et al., 2019). Secondly, if the reduced N400 effect is merely due to negative polarity, a similar effect should be seen for Aristotelian quantifiers, which included positive 'all' and negative 'none of', but this was not observed. In fact, the N400-like effect for Aristotelian quantifiers is the largest of all three classes. A second limitation is that while the theory predicts the algorithmic difference to stem from a memory component, it is not possible to ascertain whether the difference we observed is indeed related to memory. The argument made above is hypothetical: further research is needed to establish the exact cognitive and physiological nature of the observed sentence-internal verification positivity.

## 5. Conclusion

We have shown that the algorithmic verification complexity of different quantifier classes is associated with different patterns of neural responses. Our findings suggest that algorithmic aspects of language processing are subjected to the same formal constraints applicable to abstract machines. Results of previous quantifier verification experiments, to the extent that they do not take formal distinctions between quantifier classes into account, may not generalize and may not be jointly interpretable: different classes of quantifiers are provably verified using different algorithms, and thus give rise to qualitatively distinct evoked potentials. An exciting open question at the intersection of computer science and psycholinguistics is whether formal proofs about the complexity of specific computational problems, such as verification, can inform us about which class of algorithms is plausibly implemented by the brain. Our research may serve as a stepping stone in that direction and as a proof of concept for a growing literature advocating algorithmic and complexity theoretic analyses in the construction of psychological and psycholinguistic theories (Isaac et al., 2014; van Rooij & Baggio, 2020, 2021; van Rooij et al., 2019).

## Data Availability Statement

Scripts and data for this paper are available open access at DataverseNO (https://doi.org/10.18710/M6VT6Z) (Bremnes (2021)).

## Acknowledgements

JS has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement n. STG 716230 CoSaQ.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105013.

## References

Augurzky, P., Bott, O., Sternefeld, W., & Ulrich, R. (2017). Are all the triangles blue? ERP evidence for the incremental processing of German quantifier restriction. *Language and Cogntion, 9*, 603–636.

Augurzky, P., Franke, M., & Ulrich, R. (2019). Gricean Expectations in online sentence comprehension: An ERP study on the processing of scalar inferences. *Cognitive Science, 43*(8).

Augurzky, P., Hohaus, V., & Ulrich, R. (2020a). Context and complexity in incremental sentence interpretation: An ERP study on temporal quantification. *Cognitive Science, 44*(11).

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

Bach, E., Jelinek, E., Kratzer, A., & Partee, B. H. (1995). *Quantification in natural languages*. Dordrecht: Kluwer Academic Publishers. editors.

Baggio, G. (2018). *Meaning in the brain*. Cambridge, MA: MIT Press.

Baggio, G. (2020). Epistemic transfer between linguistics and neuroscience: Problems and prospects. In R. Nefdt, C. Klippi, & B. Karstens (Eds.), *The philosophy and science of language: Interdisciplinary perspectives, pages* (pp. 275–308). Cham: Palgrave Macmillan.

Baggio, G. (2021). Compositionality in a parallel architecture for language processing. *Cognitive Science, 45*(5), e12949.

Baggio, G., & Bremnes, H. S. (2017). Book review: Jakub Szymanik quantifiers and cognition. Logical and computational perspectives. Springer, 2016. *Studia Logica, 105*, 1015–1019.

Baggio, G., Stenning, K., & van Lambalgen, M. (2016). Semantics and cognition. In M. Aloni, & P. Dekker (Eds.), *The Cambridge handbook of formal semantics* (pp. 756–774). Cambridge: Cambridge University Press.

Baggio, G., van Lambalgen, M., & Hagoort, P. (2015). Logic as marr's computational level: Four case studies. *Topics in Cognitive Science, 7*(2), 287–298.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy, 4*, 159–219.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bremnes, H. S. (2021). *Data for computational complexity explains neural differences in quantifier verification*. DataverseNO. https://doi.org/10.18710/M6VT6Z.

Brodbeck, C., Gwilliams, L., & Pylkkänen, L. (2016). Language in context: MEG evidence for modality-general and -specific responses to reference resolution. *eNeuro, 3*.

Brouwer, H., & Hoeks, J. C. (2013). A time and place for language comprehension: Mapping the N400 and the P600 to a minimal cortical network. *Frontiers in Human Neuroscience, 7*, 758.

Carcassi, F., Steinert-Threlkeld, S., & Szymanik, Jakub (2021). Monotone quantifiers emerge via iterated learning. *Cognitive Science, 45*(8).

Chemla, E., Dautriche, I., Buccola, B., & El Fagot, J. (2019). Constraints on the lexicons of human languages have cognitive roots present in baboons (*Papio papio*). *PNAS, 116*, 30.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory, 2*, 113–124.

Clark, H. H. (1976). Semantics and Comprehension. *Mouton The Hague*.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*, 472–517.

Clark, H. H., & Chase, W. G. (1974). Perceptual coding strategies in the formation and verification of descriptions. *Memory and Cognition, 2*, 101–111.

Coppock, E. (2019). Quantity superlatives in Germanic, or life on the fault line between adjective and determiner. *Journal of Germanic Linguistics, 31*, 109–200.

D'Arcy, R. C. N., Connolly, J. F., & Crocker, S. F. (2000). Latency shifts in the N2b component track phonological deviations inspoken words. *Clinical Neurophysiology, 111*, 40–44.

De Santo, A., Rawski, J., Yazdani, A. M., & Drury, J. E. (2019). Quantified sentences as a window into prediction and priming: An ERP study. In E. Ronai, L. Stigliano, & Y. Sun (Eds.), *Proceedings of the fifty–fourth annual meeting of the Chicago linguistic society* (pp. 85–98).

Deschamps, I., Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition, 143*, 244–253.

Dwivedi, V. D., Phillips, N. A., Einagel, S., & Baum, S. R. (2010). The neural underpinnings of semantic ambiguity and anaphora. *Brain Research, 1311*, 93–109.

Embick, D., & Poeppel, D. (2015). Towards a computational(ist) neurobiology of language: Correlational, integrated and explanatory neurolinguistics. *Language, Cognition and Neuroscience, 30*(4), 357–366.

Faarlund, J. T., Lie, S., & Vannebo, K. I. (1997). *Norsk referansegrammatikk.* Oslo: Universitetsforlaget.

Freunberger, D., & Nieuwland, M. S. (2016). Incremental comprehension of spoken quantifier sentences: Evidence from brain potentials. *Brain Research, 1646*, 475–481.

Fritz, I., & Baggio, G. (2020). Meaning composition in minimal phrasal contexts: Distinct ERP effects of intensionality and denotation. *Language, Cognition and Neuroscience, 35*(10), 1295–1313.

Fritz, I., & Baggio, G. (2021). Neural and behavioural effects of typicality, denotation and composition in an adjective-noun combination task. *Language, Cognition and Neuroscience*, 1–23.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: most versus more than half. *Natural Language Semantics, 17*, 63–98.

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages, and computation.* Reading,Mass: Addison-Wesley.

Hubbard, R. J., Rommers, J., Jacobs, C. L., & Federmeier, K. D. (2019). Downstream behavioral and electrophysiological consequences of word prediction on recognition memory. *Frontiers in Human Neuroscience, 13*, 291.

Hunt, L., III, Politzer-Ahles, S., Gibson, L., Minai, U., & Fiorentino, R. (2013). Pragmatic inferences modulate N400 during sentence comprehension: Evidence from picture-sentence verification. *Neuroscience Letters, 534*, 246–251.

Hunter, T., & Lidz, J. (2013). Conservativity and learnability of determiners. *Journal of Semantics, 30*, 315–334.

Hunter, T., Lidz, J., Odic, D., & Wellwood, A. (2017). On how verification tasks are related to verification procedures: A reply to Kotek et al. *Natural Language Semantics, 25*, 91–107.

Isaac, A. M. C., 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* (pp. 787–824).

Jacob, S. N., & Nieder, A. (2009). Notation-independent representation of fractions in the human parietal cortex. *Journal of Neuroscience, 29*(14), 4652–4657.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge: Cambridge University Press.

Just, M. A. (1974). Comprehending quantified sentences: The relation between sentence-picture and semantic memory verification. *Cognitive Psychology, 6*, 216–236.

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior, 10*, 244–253.

Kanazawa, M. (2013). Monadic quantifiers recognized by deterministic pushdown automata. In M. Aloni, M. Franke, & F. Roelofsen (Eds.), *Proceedings of the 19th Amsterdam colloquium* (pp. 139–146).

Keenan, E., & Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy, 9*, 253–326.

Keenan, E. L., & Paperno, D. (2017). Overview. In D. Paperno, & E. L. Keenan (Eds.), *Handbook of quantifiers in natural language: Volume ii* (pp. 995–1004). Cham: Springer.

Kleene, S. C. (1951). *Representation of events in nerve nets and finite automata. Technical Report RM-704.* U.S. Air Force /RAND Corporation.

Knoeferle, P., Urbach, T. P., & Kutas, M. (2011). Comprehending how visual context influences incremental sentence processing: Insights from ERPs and picture-sentence verification. *Psychophysiology, 48*, 495–506.

Knowlton, T., Hunter, T., Odic, D., Wellwood, A., Halberda, J., Pietroski, P., & Lidz, J. (2021). Linguistic meanings as cognitive instructions. *Annals of the New York Academy of Sciences, 1500*(1), 134–144.

Kounios, J., & Holcomb, P. (1992). Structure and process in semantic memory: Evidence from event-related brain potentials and reaction times. *Journal of Experimental Psychology: General, 121*(4), 459–479.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research, 44*(1), 27–46.

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of most. *Natural Language Semantics, 19*, 227–256.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*, 177–190.

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

Matthewson, L. (2001). Quantification and the nature of crosslinguistic variation. *Natural Language Semantics, 9*, 145–189.

McMillan, C. T., Clark, R., Moore, P., Devita, C., & Grossman, M. (2005). Neural basis for generalized quantifier comprehension. *Neuropsychologia, 43*(12), 1729–1737.

McMillan, C. T., Clark, R., Moore, P., & Grossman, M. (2006). Quantifier comprehension in corticobasal degeneration. *Brain and Cognition, 62*(3), 250–260.

McMillan, C. T., Coleman, D., Clark, R., Liang, T.-W., Gross, R. G., & Grossman, M. (2013). Converging evidence for the processing costs associated with ambiguous quantifier comprehension. *Frontiers in Psychology, 4*, 153.

Mock, J., Huber, S., Bloechle, J., Bahnmueller, J., Moeller, K., & Klein, E. (2019). Processing symbolic and non-symbolic proportions: Domain-specific numerical and domain-general processes in intraparietal cortex. *Brain Research, 1714*, 133–146.

Mock, J., Huber, S., Bloechle, J., Dietrich, J. F., Bahnmueller, J., Rennig, J., Klein, E., & Moeller, K. (2018). Magnitude processing of symbolic and non-symbolic proportions: An fMRI study. *Behavioral and Brain Functions, 14*(1).

Morgan, B., Gross, R. G., Clark, R., Dreyfuss, M., Boller, A., Camp, E., Liang, T. W., Avants, B., McMillan, C. T., & Grossman, M. (2011). Some is not enough: Quantifier comprehension in corticobasal syndrome and behavioral variant frontotemporal dementia. *Neuropsychologia, 49*(13), 3532–3541.

Moschovakis, Y. N. (2006). A logical calculus of meaning and synonymy. *Linguistics and Philosophy, 29*, 27–89.

Mostowski, M. (1998). Computational semantics for monadic quantifiers. *Journal of Applied Non-Classical Logics, 8*, 107–121.

Muskens, R. (2005). Sense and the computation of reference. *Linguistics and Philosophy, 28*, 473–504.

Nieuwland, M. S. (2016). Quantification, prediction, and the online impact of sentence truth-value: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(2), 316–334.

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language, 85*(2), 203–210.

Olm, C. A., McMillan, C. T., Spotorno, N., Clark, R., & Grossman, M. (2014). The relative contributions of frontal and parietal cortex for generalized quantifier comprehension. *Frontiers in Human Neuroscience, 8*.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG,EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011*.

O'Rourke, P. L., & Van Petten, C. (2011). Morphological agreement at a distance: Dissociation between early and late components of the event-related brain potential. *Brain Research, 1392*, 62–79.

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of 'most': Semantics numerosity and psychology. *Mind and Language, 24*, 554–585.

Pietroski, P., Lidz, J., Hunter, T., Odic, D., & Halberda, J. (2011). Seeing what you mean, mostly. In J. Runner (Ed.), *Experiments at the interfaces, volume 37 of syntax and semantics* (pp. 181–217). Leiden: Brill.

Politzer-Ahles, S., Fiorentino, R., Jiang, X., & Zhou, X. (2013). Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification. *Brain Research, 1490*, 134–152.

Ratcliff, R., Sederberg, P. B., Smith, T. A., & Childers, R. (2016). A single trial analysis of EEG in recognition memory: Tracking the neural correlates of memory strength. *Neuropsychologia, 93*, 128–141.

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *TRENDS in Cognitive Sciences, 11*, 251–257.

Rugg, M. D., Mark, R. E., Walla, P., Schloerscheidt, A. M., Birch, C. S., & Allan, K. (1998). Dissociation of the neural correlates of implicit and explicit memory. *Nature, 392*, 595–598.

Sassenhagen, J., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2014). The P600-as-P3 hypothesis revisited: single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain and Language, 137*, 29–39.

Shetreet, E., Chierchia, G., & Gaab, N. (2014). When three is not some: On the pragmatics of numerals. *Journal of Cognitive Neuroscience, 26*(4), 854–863.

Spychalska, M., Kontinen, J., Noveck, I., Reimer, L., & Werning, M. (2019). When numbers are not exact: Ambiguity and prediction in the processing of sentences with bare numerals. *Journal of Experimental Psychology: Learning Memory and Cognition, 45*(7), 1177–1204.

Spychalska, M., Kontinen, J., & Werning, M. (2016). Investigating scalar implicatures in a truth-value judgement task: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience, 31*, 817–840.

Steinert-Threlkeld, S., & Szymanik, J. (2020). Learnability and semantic universals. *Semantics and Pragmatics, 12*(4).

Suppes, P. (1982). Variable-free semantics with remarks on procedural extensions. In T. W. Simon, & R. J. Scholes (Eds.), *Language, mind, and brain* (pp. 21–34). Hillsdale: Erlbaum.

Szymanik, J. (2016). *Quantifiers and cognition: Logical and computational perspectives.* Cham: Springer.

Szymanik, J., & Thorne, C. (2017). Exploring the relation between semantic complexity and quantifier distribution in large corpora. *Language Sciences, 60*, 80–93.

Szymanik, J., & Zajenkowski, M. (2010a). Comprehension of simple quantifiers: Empirical evaluation of a computational model. *Cognitive Science, 34*(3), 521–532.

Szymanik, J., & Zajenkowski, M. (2009). (2010b). Quantifiers and working memory. M. Aloni, H. Bastiaanse, T. e Jager, P. van Ormondt, & K. Schulz (Eds.). . In *Amsterdam Colloquium, 25* pp. 456–464) Berlin, Heidelberg: Springer Verlag.

Szymanik, J., & Zajenkowski, M. (2011). Contribution of working memory in parity and proportional judgments. *Belgian Journal of Linguistics, 25*, 176–194.

Talmina, N., Kochari, A., & Szymanik, J. (2017). Quantifiers and verification strategies: Connecting the dots. In A. Cremers, T. van Gessel, & F. Roelofsen (Eds.), *Proceedings of the 21st Amsterdam colloquium* (pp. 465–473).

Tichý, P. (1969). Intension in terms of turing machines. *Studia Logica, 24*, 7–21.

Tomaszewicz, B. (2011). Verification strategies for two majority quantifiers in polishI. Reich, E. Horch, & D. Pauly (Eds.). . In *Proceedings of sinn und Bedeutung, 15* pp. 597–612).

Urbach, T. P., DeLong, K. A., & Kutas, M. (2015). Quantifiers are incrementally interpreted in context, more than less. *Journal of Memory and Language, 83*, 79–96.

Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language, 63* (2), 158–179.

van Benthem, J. (1986). *Essays in logical semantics*. Netherlands: Springer.

van de Pol, I., Steinert-Threlkeld, S., & Szymanik, J. (2019). Complexity and learnability in the explanation of semantic universals of quantifiers. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual conference of the cognitive science society* (pp. 3015–3021). Montreal, QB: Cognitive Science Society.

van Lambalgen, M., & Hamm, F. (2005). *The proper treatment of events*. Malden: Blackwell.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning Memory, and Cognition, 25*, 394–417.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*, 176–190.

van Rooij, I., & Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychological Inquiry, 31*(4), 321–325.

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science, 16*(4), 682–697.

van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and intractability: A guide to classical and parameterized complexity analysis.* Cambridge: Cambridge Univeristy Press.

Vissers, C. T. W. M., Kolk, H. K. J., van de Meerendonk, N., & Chwilla, D. J. (2008). Monitoring in language perception: Evidence from ERPs in a picture-sentence matching task. *Neuropsychologia, 46*, 967–982.

Wassenaar, M., & Hagoort, P. (2007). Thematic role assignment in patients with Broca's aphasia: Sentence-picture matching electrified. *Neuropsychologia, 45*, 716–740.

Yang, H., Laforge, G., Stojanoski, B., Nichols, E. S., McRae, K., & K&rdquo;ohler, S. (2019). Late positive complex in event-related potentials tracks memory signals when they are decision relevant. *Scientific Reports, 9*, 9469.

Zajenkowski, M., & Szymanik, J. (2013). MOST intelligent people are accurate and SOME fast people are intelligent Intelligence, working memory, and semantic processing of quantifiers from a computational perspective. *Intelligence, 41*(5), 456–466.

Zajenkowski, M., Szymanik, J., & Garraffa, M. (2014). Working memory mechanism in proportional quantifier verification. *Journal of Psycholinguistic Research, 43*(6), 839–853.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162–185.

# Paper 2

## The interplay of computational complexity and memory load during quantifier verification

The final publisehd version is available in
Language, Cognition and Neuroscience 2024 ;Volum 39.(1) s. 1-23
https://doi.org/10.1080/23273798.2023.2236253

# The interplay of computational complexity and memory load during quantifier verification

Heming Strømholt Bremnes[1], Jakub Szymanik[2], and and Giosuè Baggio[1]

[1]Language Aquisition and Language Processing Lab, Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway
[2]Center for Mind/Brain Sciences and Department of Information Engineering and Computer Science, University of Trento, Italy

**Abstract**

Formal analysis of the minimal computational complexity of verification algorithms for natural language quantifiers implies that different classes of quantifiers demand the engagement of different cognitive resources for their verification. In particular, sentences containing *proportional* quantifiers, e.g., 'most', provably require a memory component, whereas *non-proportional* quantifiers, e.g., 'all', 'three', do not. In an ERP study, we tested whether previously observed differences between these classes were modulated by memory load. Participants performed a picture-sentence verification task while they had to remember a string of 2 or 4 digits to be compared to a second string at the end of a trial. Relative to non-proportional quantifiers, proportional quantifiers elicited a sentence-internal sustained negativity that was larger for 4 than for 2 digit strings. Our results suggest that the formal constraints applicable to abstract machines are of the same nature as the constraints on cognitive resources deployed during human sentence processing and verification.

***Keywords:*** Quantifiers; Computational complexity; Semantic automata; Memory; Picture-sentence verification; ERPs

## 1 Introduction

Quantification is a fundamental aspect of human cognition. It lies at the heart of our linguistic, logical, and mathematical abilities and as a consequence it has been studied extensively at least since Aristotle. In natural languages, quantitative relations are often expressed using determiners, like 'all', 'three', and 'most', that are unusually homogeneous across languages (Bach et al., 1995; Keenan and Paperno, 2017; Matthewson, 2001). Pioneering work (Barwise and Cooper, 1981; Keenan and Stavi, 1986) has demonstrated that natural language quantifiers constitute a small subset of the quantitative relations expressible with logical vocabulary. More recently, it has been shown that certain characteristic formal properties of this subset delineate learning biases for humans, non-human primates, and machine learning algorithms (Carcassi et al., 2021; Chemla et al., 2019; Hunter and Lidz, 2013; Steinert-Threlkeld and Szymanik, 2019; van de Pol et al., 2019). These findings suggest that studying natural language quantifiers can inform cognitive science about the human language capacity specifically and human cognition more generally.

In Marrian cognitive (neuro)science (Marr, 1982), information processing systems can be understood at three levels of analysis: (i) a computational level, describing a computation in terms of a function mapping inputs to outputs; (ii) an algorithmic level, detailing the stepwise procedures and subprocedures required to compute the function; and (iii) an implementational level that specifies how this algorithm is implemented in the biophysical medium of the brain. Algorithmic analyses are constrained both by the nature of the computation and by the limitations placed on the kinds of processes the brain is able to carry out. Since the algorithmic level is indispensable in mediating between the computational and the implementational levels (Baggio et al., 2015, 2016; Embick and Poeppel, 2015; Lewis and Phillips, 2015), specifying the properties of the algorithms that underlie cognitive computation is essential. It might therefore seem puzzling that algorithmic aspects of semantic processing hitherto have not received sufficient attention (Baggio, 2018, 2020). One reason for this might be the fact that meanings are notoriously hard to formalize and that such formalizations are required to study algorithms.

Natural language quantifiers are an interesting exception to this rule, because their precise meaning contributions can be formalized in *generalized quantifier theory* as relations between the cardinalities of sets (Barwise and Cooper, 1981; Peters and Westerståhl, 2006). This approach has made quantifiers a linchpin in the development of formal semantics (Partee, 2013), and it also enables the construction of verification algorithms for quantifiers, to be discussed in more detail in 1.1. Once these algorithms are specified, it is mathematically provable that quantifiers can be divided into different classes, based on the computational resources required to verify them. When determining the computational properties of quantifier verification, the difference between proportional quantifiers – e.g. 'most', 'less than half' – and other quantifiers is that proportional quantifiers cannot be verified by a simple finite-state automaton (FSA), but instead require a push-down automaton (PDA) with its memory component. In a previous study (Bremnes et al., 2022), we showed that quantifier class modulates ERP responses in a verification task: proportional quantifiers resulted in ERP effects that were absent for non-proportional quantifiers. Moreover, such effects were observed only in a verification task, and not in a task that required participants to just read and understand quantified sentences. The goal of the present study was to ascertain whether the observed differences in evoked potentials are in fact related to the usage of memory resources in the service of verification, and to gather initial evidence for the specific memory systems deployed.

## 1.1 Algorithms of quantifier verification

The idea to construct verification algorithms for natural language quantifiers originated with van Benthem (1986) and has led to many subsequent mathematical results about the computational properties of such algorithms (e.g., Kanazawa, 2013; Mostowski, 1998; Szymanik, 2016). The semantics for natural language quantifiers given in generalized quantifier theory (Barwise and Cooper, 1981; Keenan and Stavi, 1986) as (conservative and extensional) relations between cardinalities of sets, allows determiner meanings to be modeled as sets of strings of binary recognized by abstract computational models called *automata*. These are foundational tools from theoretical computer science and formal language theory, and can be used to mathematically prove differences in the minimal complexity of different computational problems (Chomsky, 1956; Hopcroft and Ullman, 1979).

The strings of binary represent the objects being quantified over as having or not having a predicated property, for example a set of circles as having the property of being red for a sentence like 'All the circles are red'. These algorithms run through all the elements in the set and for each of them check if they have that property. If, by the time a given algorithm has checked all the objects, the number of objects with the property conforms to the quantitative relation expressed by the quantifier, the sentence is true. Otherwise it is false.

Let us informally illustrate this procedure for the quantifiers 'no', 'at least four', and 'more than half', as applied to red circles. For 'no', the minimal algorithm scans all the circles, and if it does not find a red circle, the sentence is true. In the case of 'at least four', the same kind of algorithm scans all the circles and keeps track of the red circles it sees until it has reached four. At that point, all the subsequent circles are irrelevant, because the sentence will be true regardless. Both these kinds of quantifiers, so-called *Aristotelian* and *numerical* quantifiers, respectively, can be computed by the simplest kind of machine: *finite state automata* (FSA). This is not the case for 'more than half', which is a *proportional* quantifier: such quantifiers are concerned with the proportion of red to non-red circles. They provably require a memory component where an algorithm can store information about red and non-red circles, and can minimally be verified by a *pushdown automaton* (PDA). For 'more than half', the simplest algorithm keeps track of both the red circles and the non-red circles as it scans the set. Once it has scanned the final circle, it checks if the red circles outnumber the non-red circles, and if they do, the sentence is true. For formal definitions and explanations of the automata, see Szymanik (2016, chapter 4).

Importantly, this leads to two qualitatively different kinds of verification algorithms. Any algorithm for proportional quantifiers is of a different nature than the minimal verification algorithms for both Aristotelian and numerical quantifiers. It is therefore essential to distinguish between proportional and non-proportional quantifiers, because of the different computational resources required to verify them. In particular, only proportional quantifiers are predicted to require the storing and manipulation of objects in memory.

## 1.2 Previous studies

Numerous studies have examined quantifier verification (e.g., Freunberger and Nieuwland, 2016; Kounios and Holcomb, 1992; Noveck and Posada, 2003; Nieuwland, 2016; Urbach and Kutas, 2010; Urbach et al., 2015), and several have used a picture-sentence verification task for quantified sentences (Augurzky et al., 2017, 2019, 2020a,b; Hunt III et al., 2013; Politzer-Ahles et al., 2013; Spychalska et al., 2016, 2019). These studies have predominantly focused on effects of truth value and have shown that false sentences exhibit larger N400-like responses than true sentences. More interestingly, the complexity of the verification – either as a result of the picture or the sentence – manifests itself as an increased positivity after the N400 time frame and as sustained effects earlier in the sentence.

In previous experiments (Bremnes et al., 2022), we demonstrated that differences in the verification procedure for proportional quantifiers, as described above, give rise to specific ERP effects. In a picture-sentence verification task, participants saw red and yellow circles and triangles and had to judge the truth value of quantified sentences, e.g., ‘All the circles are red’. In addition to the expected N400-like effects of truth value at the final word, and to a post-N400 positivity for proportional quantifiers, we observed a sustained positivity in the P600 time-window on the completion of the subject noun phrase (‘Most of the circles’) for proportional quantifiers compared to non-proportional. This pattern was also observed in the only other study that has explored ERP effects of quantifier class (De Santo et al., 2019).

The literature on memory and quantifier verification has hitherto been disjoint, but the nature of the present project necessitates their integration. It is therefore pertinent to discuss different ERP components that have been associated with various kinds of memory, as well as their functional interpretation, in order to make more refined predictions about which components could plausibly be modulated in a verification task.

Late positivities, such as the one found in Bremnes et al. (2022), have often been described in the literature on recollection memory, where they are labelled the *late positive component* (LPC) or the *parietal old/new effect* (e.g., see Rugg et al., 1998; Ratcliff et al., 2016; Hubbard et al., 2019; Yang et al., 2019). This effect is observed when participants are recalling contextual details of a stimulus (Rugg and Curran, 2007), when recollection is task relevant (Yang et al., 2019). Positive slow waves have also been observed in paradigms that examined short-term or working memory (for discussion see Baddeley, 2012, and references therein), such as serial recall tasks (Kusak et al., 2000), delayed matched to sample (DMTS) tasks (McEvoy et al., 1998; Ruchkin et al., 1992), the Sternberg task (Pelosi et al., 1992, 1995, 1998), or other digit span tasks (Lefebvre et al., 2005; Marchand et al., 2006), and have been argued to index retrieval of information from short-term memory (García-Larrea and Cézanne-Bert, 1998).

However, sustained negative ERPs have also been reported for increased memory load. The sustained anterior negativity (SAN) has been reported in sentence processing when working memory resources have to be recruited for the recomputation of discourse models (Baggio et al., 2008; Müller et al., 1997; Münte et al., 1998) or as a result of referential ambiguity in the model (van Berkum et al., 1999, 2003). Sustained negativities have also been shown to arise under increased working memory load in sentence processing (Vos et al., 2001) or other working memory tasks, for instance during the retention interval of DMTS tasks (Ruchkin et al., 2003) and in visual working memory tasks (Axel and Müller, 1996; Rösler et al., 1997; Ruchkin et al., 1990, 1992; Vogel and Machizawa, 2004). These effects are similar in distribution to the left anterior negativity (LAN), occasionally accompanied, in biphasic patterns, by P600 effects in morphosyntactic violation paradigms (Baggio, 2008). However, studies have reported both short-lived and sustained left anterior negative ERPs. It is not clear whether short-lived LAN effects index working memory load in sentence processing (Fiebach et al., 2001; King and Kutas, 1995; Kluender and Kutas, 1995; Vos et al., 2001). Sustained left-anterior negativities seem more likely candidates ERP signatures of working memory usage during sentence processing.

Interestingly, what presents itself as a posterior negative slow wave in adults is observed as an anterior positivity in children (Barriga-Paulino et al., 2014), a reminder that the same underlying process may manifest itself in different polarities depending on brain anatomy and the orientation of dipole generators (for discussion, see Luck, 2014). This can also be seen in the differing polarities of slow waves over posterior and frontal regions in certain working memory paradigms, such as the n-back task (Bailey et al., 2016; McEvoy et al., 1998) and DMTS (Ruchkin et al., 1990, 1992). Furthermore, scores from working memory assessments have been shown to be correlated with sustained effects (Adam et al., 2020; Amico et al., 2015; Barriga-Paulino et al., 2014; Fukuda et al., 2015; Harker and Connolly, 2007; Lefebvre et al., 2013; Luria et al., 2016; Marchand et al., 2006). However, while some studies have found a larger ERP effect to be associated with higher performance, others have found the reverse pattern, i.e., worse performance associated with a larger effect. In language processing, larger sustained negativities have been associated with lower reading span scores when dividing participants into high and low span groups (Fiebach et al., 2002; Vos et al., 2001). A reduction of the P400 for 5 versus 1 digits in the Sternberg short-term memory task has also been shown to correlate with better task performance (Pelosi et al., 1992). By contrast, an increase in the LPC is associated with higher accuracy in recognition memory paradigms (Harker and Connolly, 2007), increased SAN amplitudes have been associated with greater auditory short-term memory capacity (Lefebvre et al., 2013), and a more negative parietal slow wave is associated with higher scores on working memory tests in the visual working memory literature (Barriga-Paulino et al., 2014; Luria et al., 2016). This demonstrates that such ERPs are modulated by individual working memory capacity, but that the direction of the modulation might depend on the task or on the specific memory systems involved.

## 1.3 The present study

The aims of the present study were to determine (1) whether the ERP differences between proportional and non-proportional quantifiers first reported in Bremnes et al. (2022) are replicable, and (2) whether these differences are related to memory, as predicted by the automata theory. To that end, we conducted an EEG experiment using the same picture-sentence verification task as our previous study, augmented with a digit matching task that allowed us to manipulate memory load. Before each trial, participants saw a string of 2 or 4 digits that they had to remember while completing the verification task. Once the verification task was completed, they saw another string of digits that either matched the original string or differed by a single digit, and had to decide whether the two strings were the same or different. In addition, participants performed a series of preliminary tasks that allowed us to test whether the electrophysiological differences were related to individual differences in working memory, attention, and control capacities. Negative proportional quantifiers have been associated with some of the effects observed in our previous study. Here, we decided to increase the number of trials for positive and negative proportionals compared to Bremnes et al. (2022), so that we would be able to rule out the possibility that negative proportionals are driving the effect. A more detailed description of the task is found in 2.1 below.

Regarding memory load, two results would corroborate the theory. Firstly, memory load, introduced by the digit span task, could increase processing differences between the quantifier classes, resulting in larger amplitude differences between proportional and non-proportional quantifiers. In this case, memory load from verification and digit matching may affect the proportional quantifiers more because it strains working memory capacity. Alternatively, memory load could attenuate the differences between quantifier classes, resulting in smaller differences between them. This pattern could be explained by finite memory: memory capacity may already be at ceiling with proportional quantifiers, but not with non-proportional quantifiers. In both scenarios, memory would affect the two quantifier classes differently, so both outcomes would support the conclusion that the verification differences are related to memory.

However, there are two additional outcomes worth considering. The memory load from the verification task and from the digit matching task could result in an additive effect, impacting proportional and non-proportional quantifiers equally: the difference between the two quantifier classes would then be similar between memory loads. Although strictly compatible with the theory, this result would be inconclusive because, in that event, it is conceivable that the difference is related to something other than memory. Finally, it is possible that memory load does not affect brain responses at all, namely that there is no difference between the high and the low memory condition. This is more problematic for the theory, since this would imply that the differences are not related to memory at all.

On the basis of previously observed behavioral effects (Zajenkowski et al., 2011; Zajenkowski and Szymanik, 2013; Zajenkowski et al., 2014), we expect individual differences in the preliminary tasks to correlate with the ERPs. However, the direction of this correlation is not predicted, as working memory capacity and amplitude have displayed both positive and negative correlations in the past (see above). The fact that some people are faster or more accurate in these tasks need not impact the verification process itself. This issue is particularly important, considering the fact that the automata theory does not predict the involvement of specific memory systems or their associated effects. The relevant automata theoretic notion of memory is abstract, and it is an empirical question, partially considered here, which human memory systems are involved. Relatedly, while the complexity analyses presented here remain on the computational level, a growing body of work attempt the exact specification of verification algorithms for natural language quantifiers (Hackl, 2009; Hunter et al., 2017; Knowlton et al., 2021; Lidz et al., 2011; Pietroski et al., 2009, 2011; Talmina et al., 2017; Tomaszewicz, 2011). In this literature, truth conditionally equivalent quantifiers are shown to be verified differently on the basis of whether they benefit from certain properties of the visual stimulus, such as grouping effects, or not. From this finding, scholars infer that these quantifiers recruit different non-linguistic systems – such as cardinality estimation based on the approximate number system or exact counting (see e.g. Dehaene, 2011; Odic and Starr, 2018), or one-to-one mapping (e.g. Feigenson, 2005) – depending on what appears to be their canonical verification procedure. However, rather than trying to detect differences within quantifier classes, what we are trying to demonstrate is that, irrespective of the specific algorithms implemented by the brain, at the very least proportional quantifier verification involves memory resources of some kind, that verification of non-proportional quantifiers do not.

# 2 Methods

## 2.1 Design

The study used a 2×2×2 design with the factors Quantifier Class (Proportional/Non-proportional), Digit Load (2/4), and Truth-Value (True/False). Each trial consisted of two tasks: after reading the sentence, the participant had to perform a sentence-picture verification task; next, they had to recall a string of 2 or 4 digits presented at the start of the trial and decide whether it was the same or different from another string of digits presented at the end of the trial. The set-up was comparable to that of our previous study (Bremnes et al., 2022). Specifically, the picture was presented before the sentence to avoid eye-movement disturbances of the EEG signal. Furthermore, the same picture was presented before each trial in a block. Participants had the opportunity to study this picture for as long as they wanted at the beginning of each block. This was (i) because remembering the picture is a prerequisite for performing the task, and we wanted to make sure that participants could memorize the picture, and (ii) because we did not want memory encoding or recall of the picture to interfere with the deployment of memory resources relevant to verification or digit recall. A potential worry is that all quantifier classes require some form of memory in this set-up. However, as noted above, the automata theory shows that proportional quantifiers require additional memory resources to maintain and compare two sets of objects in memory, which is predicted to increase memory load only for this class of quantifiers (Bremnes et al., 2022). This set-up ensures a stable baseline, where the differences detected are plausibly related to the experimental manipulations, and not to differences in encoding or recollection of the picture.

## 2.2 Participants

Fifty native speakers of Norwegian (28 female; mean age 22.98, sd = 2.93; age range 19-30), with normal or corrected to normal vision and no psychiatric or neurological disorders, were recruited from the local student community. Two of these did not meet the inclusion criteria of having an average of at least 80% artifact free trials per condition, and were excluded from the final data analysis. We then analyzed data from 48 participants (26 female; mean age 22.95, sd = 2.9; age range 19-30). All participants gave their written informed consent and were compensated with a voucher. The study had been approved prior to commencement by the Norwegian Centre for Research Data (NSD; project nr. 455334).

## 2.3 Materials and tasks

At the beginning of a session, participants were administered three tests of executive function, memory, and attention. All tests began with a series of practice trials (10 for the Eriksen task, 5 for the Sternberg task, 4 for the Brown-Peterson task) before the main experiment began (details below).

The first task was a version of the classic Eriksen flanker task (Eriksen and Eriksen, 1966), aimed at measuring attention. Participants were shown rows of arrows and had to determine in which direction the middle arrow pointed. The rows could be either congruent (all arrows pointed in the same direction) or incongruent (different directions). Each participant saw 60 rows (30 congruent) with an equal number of correct right and left responses.

In order to test working memory capacity, the second task implemented a Sternberg scanning paradigm (Sternberg, 1966), in which participants saw 4, 6, or 8 digits presented consecutively. They then saw a digit in red and had to determine whether this digit was also included in the preceding digit sequence. Each sequence length was presented 16 times, with 8 trials where the target number was presented and 8 trials where it was absent.

The third task was a Brown-Peterson short-term memory task (Brown, 1958; Peterson and Peterson, 1959), targeted at working memory capacity in the presence of distractors. Each trial consisted of a to-be-remembered consonant trigram (e.g. 'FCQ') and a number between 150 and 500, from which the participant had to count backwards in threes out loud. The counting lasted 4, 6, or 12 seconds, and the participant was subsequently prompted to recall the trigram or, as a control trial, the latest number they counted. There were 8 trials for each counting interval, or 24 trials in total, with 3 controls for each interval length. We opted for 4, 6, and 12 as a short, medium, and long condition respectively, which is comparable to intervals used previously (Neath et al., 2019; Quinlan et al., 2015). These particular intervals allowed us to keep the task manageable in terms of total duration. It has been shown that accuracy in this task decreases sharply from 1 to 9 seconds but flattens out after that, so that there is only a small accuracy difference between, e.g., 12 and 18 seconds (Rai and Harris, 2013).

As mentioned in 2.1 above, the main task was to memorize a string of 2 or 4 digits, then perform a picture-sentence verification task, and finally judge whether another string of digits matched the string seen at the beginning of each trial.

For the digit matching task, we opted for one high and one low digit load conditions. Previous studies (Szymanik and Zajenkowski, 2010, 2011) found that, with 4 and 6 digits, digit recall was poor at 6 digits. In contrast, performance on the verification task increased, both in terms of accuracy and RT, for 6 digits compared to 4, suggesting that the task was too difficult with 6 digits. We therefore used 2 digits as the low load condition and 4 digits as the high load condition. First, we constructed random strings of 2 and 4 digits. For half of these, we also created mismatch strings by replacing one random digit in each string with another random digit. For example, if the string was 4459, we would replace the second digit with 8 to create 4859, or the third digit with 2 to create 4429. The decision to make digit string pairs minimally distinguishable by a single digit was made because, with completely different strings, participants could easily adopt a strategy where they only memorized the first two digits and still be correct in many cases. This would effectively render the distinction between 2 and 4 digits useless.

For the verification task, we constructed 8 pictures consisting of clustered red and yellow circles and triangles in a 2×2 grid. The grid location, number and color of these shapes were varied pseudorandomly. The grid design with a 2×2 potential shape by color alternation secured that participants could not know the truth-value of the sentence before reading the final word. The number of objects at each grid location ranged from 2 to 5. For every picture in which the shapes of one type (e.g., circles) were all in one color, the other was always in different colors. Each picture was shown for all trials in one block, meaning that there were 8 blocks in the experiment. See *Supplementary material A*, section I, for all pictures.

The sentences were simple subject-predicate copular sentence, in which a certain color was predicated

| Quantifier Class | Quantifier | Shape | Copula | Color |
|---|---|---|---|---|
| Aristotelian | Samtlige av | sirklene<br>*the*<br>*circles* | er<br>*are* | røde<br>*red* |
| | Ingen av | | | |
| | Enkelte av | | | |
| Numerical | Tre av | | | |
| | Fire av | | | |
| | Fem av | | | |
| Positive Proportional | Flesteparten av | trekantene<br>*the*<br>*triangles* | | gule<br>*yellow* |
| | Flest av | | | |
| | Over halvparten av | | | |
| Negative Proportional | Minsteparten av | | | |
| | Færrest av | | | |
| | Under halvparten av | | | |

Table 1: The experimental sentences were constructed by combining every element of one column with every element of the other columns, resulting in 12 × 2 × 1 × 2 = 48 different sentences. For the translations of the quantifier column, see main text. All experimental sentences with translations can be found in *Supplementary material A*, section II.

of a certain quantity of shapes s (e.g. 'Flest av sirklene er røde', *Most of the circles are red*). We wanted the syntax and the semantics of the sentences to be as closely matched as possible, aside from the quantifier manipulation. We therefore decided to only use quantifiers in partitive constructions, which is the most natural – and, for some quantifiers, the only – way to express quantitative relations between definite objects in Norwegian. This also ensured that all shape nouns were definite plurals and that adjectives agreed in number with these shape nouns. We used 12 quantifiers, 3 of each type. The non-proportional quantifiers were Aristotelian ('samtlige av': *all of*; 'ingen av': *none of*; 'enkelte av': *some of*) and numerical quantifiers ('tre av': *three of* ; 'fire av': *four of*; 'fem av': *five of*). The proportional quantifiers included three positive ('flesteparten av': *the majority of*; 'flest av': *most of*; 'over halvparten av': *more than half of*) and three negative quantifiers ('minsteparten av': *the minority of*; 'færrest av': *fewest of*; 'under halvparten av': *less than half of*). Combined with two shape nouns and two color adjectives, this yields a total of 48 experimental items (Table 1). Note that Norwegian and English differ with regards to the definiteness of proportional quantifiers (Coppock, 2019). See *Supplementary material A*, section II, for all experimental sentences with translations.

Each sentence was presented once for every truth-value and digit load: each sentence was true twice, once with 2 digits and once with 4 digits, and false twice, once for each digit condition. Thus, there were 192 trials overall, with 96 true/false trials in the main verification task and 96 trials with 2/4 digits in the digit matching task. There were 48 trials in each cell in the 2×2×2 design. This number is standard in ERP research, but this meant that there were only 12 trials per quantifier type (e.g., Aristotelian) by digit load by truth-value: it was then acknowledged that it would not be possible to compare truth-value by digit load EEG effects at the level of each individual quantifier type.

As mentioned, the 8 pictures constituted the block structure, and consequently there were 24 trials in each block. Because the picture remained the same within a block and there were more possible quantifier by truth-value by digit load triplets than pictures (16 triplets per quantifier), not all sentences were shown after a particular picture and some sentences had to be shown twice within the same block, that is, both digit conditions in one block. However, both truth-value and digit load were evenly balanced both within each block (12 true/false, 12 2/4 digits) and overall. It was not possible to match the number of 2 and 4 digit matches within a block (range of 2/4 digit matches: 5-7) while simultaneously retaining the balance overall. Note that this cannot possibly affect the EEG, as participants have no way of knowing whether the upcoming digits will match or mismatch the memorized string when the EEG is recorded, i.e., when they read the sentence. To avoid conflicting interpretations, quantifiers that give rise to scalar implicatures, i.e., the inferred negation of a stronger meaning (see e.g. Horn, 1972; Levinson, 1983, 2000), were not shown in contexts where both the semantic and pragmatic meanings are available. First, 'enkelte av' (*some of*), which gives rise to a scalar implicature *not all*, was not shown in pictures where the denotation of the shape noun was all in one color, e.g. 'Some of the circles are red', when there were only red circles. For the same reason, we also avoided proportional quantifiers in such contexts, e.g. 'more/less than half of the triangles are red', when all the triangles had the same color. Second, numerical quantifiers, that can have both an *exactly* and an *at least* interpretation, were never shown after pictures where the number of shapes in the predicated color exceeded the number denoted by the quantifier, e.g., 'three of the circles are yellow', when there were four yellow circles. Finally, if one shape was all in one color and the sum of the shapes in the two grid locations matched the number denoted by a numerical quantifier, e.g., if there were $2 + 3 = 5$ yellow triangles, then sentences containing that quantifier were not shown.

Trials were randomized within each block. To counterbalance sentence types within a block, we also constructed 2 randomized orders of the blocks, that we ran both forward and backward for a total of 4 different randomizations, so that participants would encounter the sentence types at different stages of the experiment.

## 2.4 Procedure

Each experimental session began with participants signing their informed consent sheet. They were then instructed about the three preliminary tests described in section 2.3, before they were seated in front of an LCD computer screen in a dimly lit, sound attenuated, and electrically shielded EEG booth. The same booth was used for the three preliminary tests, administered without EEG, and for the main experiment. Participants then performed the three tests in order: Eriksen flanker task, Sternberg scanning, and Brown-Peterson short-term memory task. Each test began with an on-screen reminder of the instructions, as well as practice trials. After they had completed these tests, participants were prepared for EEG recording, as described in 2.5 below. While the electrodes were mounted, participants received instructions about the task: they were told that they had to judge whether each sentence was true of the preceding picture, using two predefined response buttons, while at the same time remembering a string of 2 or 4 digits, and that after the truth-value judgement they would have to assess whether another string of digits matched the original string by using the same response keys as in the verification task. They were told to respond as soon as they knew the answer, but that accuracy was more important than speed. The truth values coded by the different response keys were counterbalanced between blocks. Which key corresponded to true or false was indicated by two squares with the words 'sant' (*true*) or 'usant' (*false*) on horizontally opposing sides of the screen, whose left-right order mirrored the relative keyboard position of the response

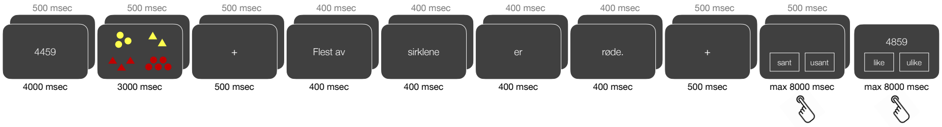


Figure 1: Structure of a single trial.

keys. This information was provided at the beginning of each block and every time they had to respond. Finally, they were instructed not to blink or move while they read the sentences, and that if such activities were necessary, they should only take place when looking at the picture or when they saw a fixation cross.

Each block began with the following preamble: participants were first informed about which buttons corresponded to true and false; they were then presented with the picture that would also be shown in each trial in the block, advised to study this picture carefully, and told to press either response button to begin with the trials. There was no time limit on how long they could study the picture. Each trial began with the presentation of a string of 2 or 4 digits for 4 seconds, preceded and followed by 500 msec of blank screen and a 500 msec fixation cross. Next, the picture was presented for 3 seconds, before another identically timed blank-screen fixation-cross pair. The sentence was presented visually in 4 chunks, where the first chunk contained the quantifier (2-3 words) and each of the remaining three contained only a single word (noun, copula, and adjective) (see Table 1, where each column represents one chunk). The reason the quantifier was presented in one chunk, was to ensure that all trials were of the same length, which is a prerequisite for comparing the different stages of the verification processes. Each chunk was shown for 400 msec with a 400 msec onset delay. Following this sequence was another 500 msec blank screen and a 500 msec fixation cross, before the response key indicators reappeared on the screen and participants could judge whether the sentence was true or false. When participants responded, or if they had not responded before 8 sec had passed, another blank screen and fixation cross pair preceded the response screen for the digit task. This screen contained the response key information, except the words for true and false were replaced by 'like' (*same*) and 'ulike' (*different*) together with the second string of numbers in the center of the screen. When participants had responded, or another 8 second time limit had expired, another identical trial started immediately (See Figure 1 for an example trial). After all 24 trials in a block had been completed, the experiment was paused and the participants were free to choose the duration of the break. The next block began when the participant pressed either response button. Each experimental session usually lasted between 2 and 2:30 hours, including the preliminary tests (20-25 mins), EEG setup (30-40 mins), and the main experiment with breaks (1:10-1:30 hours).

## 2.5 EEG-recording

EEG signals were recorded from 32 active scalp electrodes (Fp1, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, TP9, CP5, CP1, CP2, CP6, TP10, P7, P3, Pz, P4, P8, PO9, O1, Oz, O2, and PO10), using the actiCAP system by Brain Products GmbH. The implicit reference was placed on the left mastoid, and all channels were re-referenced off-line to the average of signals from the mastoids using TP10 on the right mastoid. EEG data were sampled at 1000 Hz using a 1000 Hz high cutoff filter and a

10 sec time constant. Impedance was kept below 1 kOhm across all channels throughout the experiment.

## 2.6 Data Analysis

Accuracy and reaction time data were collected for both the sentence verification and the digit recollection tasks, also to be able to compare our results with those of previous behavioral studies. However, accuracy was further used to ensure that participants were performing the task correctly. Note that reaction times are not a valid measure of the difficulty of the verification procedure, as participants could not respond as soon as they knew the answer, when the final word was presented, but had to wait for the response buttons to appear on screen 1400 msec later. For digit matching, this was not an issue, since participants could judge whether the post-trial numbers matched the pre-trial numbers immediately upon their presentation. Missed trials, where participants took too long to respond, were excluded from the analysis.

EEG data were analyzed using FieldTrip (Oostenveld et al., 2011). 1000 msec epochs, with a 200 msec pre-stimulus baseline, were extracted at the quantifier, at the noun, and at the sentence-final adjective. Trials with voltage values exceeding $\pm 150\mu V$ relative to baseline in one or more electrodes were excluded. Trials contaminated by eye movements were also excluded by thresholding the z-transformed value of the preprocessed raw data from Fp1 and Fp2 in the 1–15 Hz range. The remaining trials were subjected to a 30 Hz low-pass filter. ERPs were computed by averaging over all trials in each condition for individual participants, before sample-level ERPs were computed by averaging across participants.

ERPs were analyzed using non-parametric cluster-based statistics (Maris and Oostenveld, 2007), using the default alpha thresholds (.05) at both the sample and cluster levels. To assess ERP differences between two conditions, each sample (i.e., channel-time pair) was compared by means of a $t$-test. Adjacent samples passing a test would be added to form a cluster, and their $t$-values were summed. To determine whether two conditions were significantly different, $p$-values were estimated by using Monte Carlo simulations. For each cluster, all participant level channel-time pairs were collected into a single set before being randomly partitioned into two subsets. This procedure was repeated 1000 times. The cluster-level $p$-value was the number of random partitions that had a larger test statistic than the observed data. The output here is a (possibly empty) set of spatio-temporal clusters in which two conditions differ: we report the $T_{sum}$ in each cluster, cluster size (S), and estimated $p$-values for the highest ranked clusters.

To assess interaction effects between Quantifier Class and Digit Load, we extracted participant-level amplitudes for all channel-time pairs in the relevant clusters and we used participant mean amplitude as the dependent variable in a mixed-effect linear regression with Quantifier Class, Digit Load, and their interaction as independent variables. To determine whether working memory, attention and executive function scores were related to the ERP data, z-transformed overall accuracy ($z = \frac{x-m}{sd}$) for the Sternberg and Brown-Peterson tasks, and z-transformed median reaction time difference between congruent and incongruent trials in the Eriksen flanker task, as well as their interaction with Quantifier Class and Digit Load, were also included in the model. The models had random intercepts by participant and were estimated using the lmer function of the lme4 package (Bates et al., 2015) in R, and $p$-values were computed using the lmerTest package (Kuznetsova et al., 2017). We also computed individual level $T_{sum}$s in relevant clusters, and constructed models with these as the dependent variable, instead of mean

| | Overall | | | |
|---|---|---|---|---|
| | Accuracy | | RT | |
| | M | SD | M | SD |
| Proportional | 0.926 | 0.263 | 1748.3 | 1297.5 |
| Non-Proportional | 0.910 | 0.287 | 1531.2 | 1055.9 |

| | 2 digits | | | | 4 digits | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | | RT | | Accuracy | | RT | |
| | M | SD | M | SD | M | SD | M | SD |
| Proportional | 0.925 | 0.263 | 1724.8 | 1281.9 | 0.926 | 0.263 | 1772.9 | 1313.5 |
| Non-Proportional | 0.905 | 0.294 | 1554.8 | 1099.4 | 0.915 | 0.280 | 1509.1 | 1013.4 |

Table 2: Descriptive statistics for the linguistic verification task by Quantifier Class, with means and standard deviations of accuracy and reaction time overall and in the two Digit conditions.

| | Overall | | | |
|---|---|---|---|---|
| | Accuracy | | RT | |
| | M | SD | M | SD |
| 2 Digits | 0.915 | 0.279 | 1503.7 | 981.9 |
| 4 Digits | 0.888 | 0.315 | 1730.5 | 1005.4 |

| | Proportional | | | | Non-Proportional | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | | RT | | Accuracy | | RT | |
| | M | SD | M | SD | M | SD | M | SD |
| 2 Digits | 0.914 | 0.280 | 1488.6 | 971.1 | 0.916 | 0.277 | 1518.4 | 992.4 |
| 4 Digits | 0.894 | 0.308 | 1703.7 | 955.1 | 0.882 | 0.323 | 1756.8 | 1014.7 |

Table 3: Descriptive statistics for the digit matching task by number of digits, with means and standard deviations of accuracy and reaction time overall and in the two Quantifier Class conditions.

amplitude (Marchand et al., 2002, 2006).

# 3 Results

## 3.1 Behavioral results

In the sentence verification task, accuracy was high in all conditions, regardless of quantifier class or how many digits needed to be stored in memory (Table 2). Reaction times were markedly longer than in our previous experiment, which did not involve a digit span task. As in our previous study, however, standard deviations for reaction time data were large. Recall that the response is not produced immediately upon knowing the truth value, but after 1400 msec, when the response screen is displayed. The main function of the behavioral data was to ensure that participants were correctly performing the task, and the results confirm that they were. The reader is referred to *Supplementary material B*, section A, for inferential statistics.

Turning to the digit task, we also found very high accuracy overall and for each digit condition (Table 3). Response times were on average longer for 4 digits than for 2 digits, and, contrary to response times for the sentence verification task, there is reason to believe that response times here are representative of the underlying memory process, since there was no delay between the task and the response.

Turning lastly to the results of the three preliminary tests, means and standard deviations are found

| | M | SD |
|---|---|---|
| Eriksen | 62.250 | 32.356 |
| Sternberg | 0.866 | 0.072 |
| Brown-Peterson | 0.383 | 0.182 |

Table 4: Descriptive statistics of the measures of executive function. The measure for the Eriksen task is the difference in median reaction time for congruent and incongruent trials in msec. For the Sternberg and the Brown-Peterson, the measure is overall accuracy.

| | DAcc | DRT | QPAcc | QNPAcc | Eriksen | Sternberg | BP |
|---|---|---|---|---|---|---|---|
| DAcc | 1 | | | | | | |
| DRT | -0.162 | 1 | | | | | |
| QPAcc | **_0.691***_** | -0.354* | 1 | | | | |
| QNPAcc | **0.443**** | -0.345* | **_0.561***_** | 1 | | | |
| Eriksen | 0.006 | 0.115 | -0.223 | -0.263 | 1 | | |
| Sternberg | **0.397**** | -0.249 | 0.365* | 0.361* | -0.342* | 1 | |
| BP | **0.394**** | -0.097 | 0.290* | **_0.490***_** | -0.239 | 0.324* | 1 |

Table 5: Correlation matrix of behavioral and working memory measures, where DAcc = Digit Accuracy, DRT = Digit RT, QPAcc = Proportional quantifier accuracy, QNPAcc = Non-Proportional quantifier accuracy, BP = Brown-Peterson task. Pearson correlation coefficients are reported with coded significance values: * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$. After Bonferroni correction ($p < 0.007$), only the correlation between DAcc and the other variables, between QPAcc and QNPAcc, and between QNPAcc and BP was significant.

in Table 4. Of particular note is that accuracy in the Sternberg task is very high and exhibits very little variance, while accuracy in the Brown-Peterson task is quite low.

We found strong correlations of accuracy in the digit matching task with the verification task and the Sternberg and Brown-Peterson tasks (Table 5). The correlation is stronger for Proportional than for Non-Proportional quantifiers. There is also a strong correlation between accuracy in the verification task for Proportional and Non-Proportional quantifiers. The Brown-Peterson score is most strongly correlated with verification accuracy for Non-Proportional quantifiers.

## 3.2 ERP results

### 3.2.1 Sentence-final effects: Adjective

We began by analyzing the effects on the sentence-final adjective: the earliest point in the sentence where its truth value could be known. The waveforms (Figure 2) display a similar pattern to that found in the previous study: True and False sentences diverge after the N200, with False trials displaying a continuous negative-going deflection that overlaps temporally with the P300 wave in True trials. The two truth values largely reconverge around 450 msec. This waveform difference is also reflected in the statistics (Figure 2): we see a broadly distributed negative effect of False versus True (first-ranked negative cluster, NEG1: $T_{sum} = -16685.102$, $S = 3629$, $p = 0.001$). The cluster begins at around 250 msec and ends at around 420 msec after the onset of the adjective, with the broadest distribution and largest difference between 310 and 380 msec and the peak around 350 msec. The effect is largest on centro-parietal electrodes.

Next, we consider the effect of Digit Load. Visual inspection of the ERPs reveals that 4 and 2 Digit trials diverge around the P300 (Figure 3). From this point onward, the 4 Digit trials are distinctly more positive than the 2 Digit trials, at least up until 500 msec. This effect is confirmed by statistical analysis

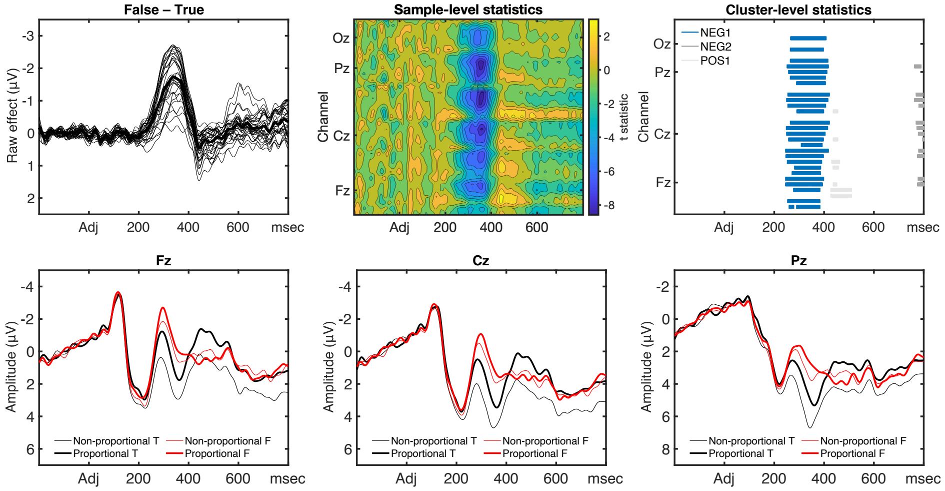


Figure 2: ERP effects of truth value (False – True) across quantifier classes (upper row), time locked to the onset of the sentence-final adjective (0 msec). Raw effect waveforms (upper left) are displayed along with contour maps of sample-level statistics (upper middle) and raster plots of cluster-level statistics (upper right). Clusters with an associated $p$-value below the specified threshold ($\alpha = 0.05$) are shown in blue shades; all other clusters (gray shades) were statistically not significant. ERP waveforms of midline electrodes (bottom row), time locked to the onset of the sentence-final adjective (0 msec). True trials are shown in black, False trials in red. Proportional quantifiers in thick lines, and Non-Proportional in thin.

(Figure 3). We found a positive cluster (first-ranked positive cluster, POS1: $T_{sum} = 2356.829$, $S = 929$, $p = 0.049$) with a central, but more posterior distribution at 260-340 msec.

The last main effect we consider, is the effect of Quantifier Class. This manipulation appears to have a similar effect on the waveforms as the Truth Value manipulation. Proportional Quantifiers diverge from Non-Proportional after the N200, where the negative ERP shift is greater for Proportional than Non-Proportional (Figure 4). Statistical analyses reveal a broadly distributed negative cluster (first-ranked negative cluster, NEG1: $T_{sum} = -6943.639$, $S = 2260$, $p = 0.015$) around 260 to 410 msec after adjective onset, and a smaller cluster (NEG2: $T_{sum} = -1797.026$, $S = 719$, $p = 0.079$) from 500 to 570 msec (Figure 4).

To sum up the main effects, there are clear effects of Truth Value, Quantifier Class, and Digit Load. False trials and Proportional quantifiers are both associated with a more negative going deflection in the 250-400 msec range, compared to their True and Non-Proportional counterparts. By contrast, 4 Digits is associated with a more positive going deflection than 2 Digits in approximately the same time window.

In addition, we examined the contrast between Proportional and Non-proportional quantifiers for 4 Digit trials and 2 Digit trials separately, on the assumption that working memory load would interact with memory usage for quantifier verification. We found that the negativity for Proportional quantifiers is driven by the effect in the 4 Digit condition (Figure 4): there were large and almost adjacent negative clusters between approximately 160 msec and the end of the epoch (NEG1: $T_{sum} = -12537.21$, $S = 4294$, $p = 0.002$; NEG2: $T_{sum} = -10599.67$, $S = 3960$, $p = 0.004$), which were not found in the 2 Digit condition (no significant clusters). We also compared positive and negative Proportional Quantifiers to make sure

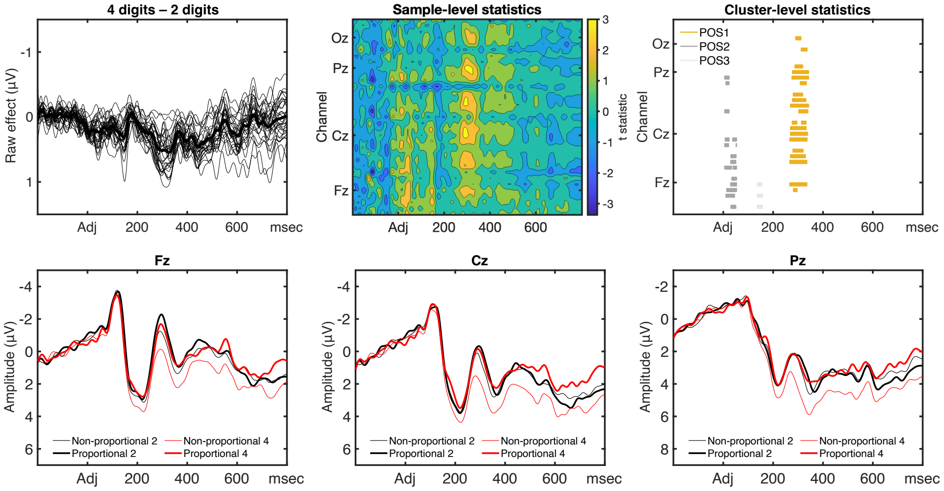


Figure 3: ERP effects of Digit Load (2 Digits – 4 Digits) across quantifier classes (upper row), time locked to the onset of the sentence-final adjective (0 msec). Raw effect waveforms (upper left) are displayed along with contour maps of sample-level statistics (upper middle) and raster plots of cluster-level statistics (upper right). Clusters with an associated $p$-value below the specified threshold ($\alpha = 0.05$) are shown in yellow shades; all other clusters (gray shades) were statistically not significant. ERP waveforms of midline electrodes (bottom row), time locked to the onset of the sentence-final adjective (0 msec). 2 Digit trials are shown in black, 4 Digit trials in red. Proportional quantifiers in thick lines, and Non-Proportional in thin.

that the effects of proportionality were not caused exclusively by the negative quantifiers. We found no significant differences overall, nor for any Digit Load or Truth Value comparison.

The results from the sentence final adjective suggest two conclusions. Firstly, there are clear effects of Truth Value, comparable to those found in our previous study, suggesting that at the time of adjective onset, participants know whether the sentence is True or False. Secondly, these Truth Value effects may be modulated by Quantifier Class and Digit Load. Indeed, most of the differences are found in the Truth Value effect time window (i.e., 250-400 msec), which is compatible with an effect of Quantifier Class and Digit Load on verification. However, these results cannot be attributed to modulations of a single ERP component, as the differences that reach significance in the different comparisons originate at different points in the epoch.

### 3.2.2 Sentence internal effects: Noun

Because a truth value has been computed at the sentence final adjective, as evidenced by the truth value effects we observe, a verification procedure is plausibly completed by this point. Consequently, we expect the effects of memory storage on the verification algorithm to occur earlier in the sentence, i.e., at the noun, as was the case in our previous study. Because the truth value could not be known at this point in the sentence, we did not distinguish between true and false trials.

We first examined the overall effect of Quantifier Class, comparing Proportional to Non-Proportional quantifiers irrespective of Digit Load. Upon visual inspection, ERP differences seem to occur early in the

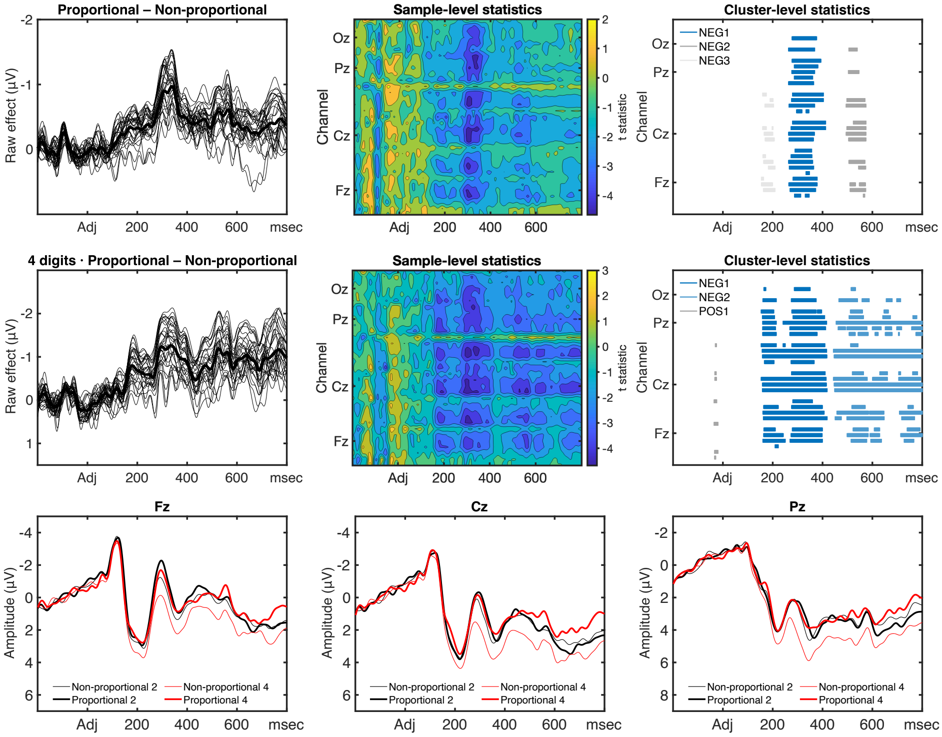


Figure 4: ERP effects of Quantifier Class (Proportional – Non-Proportional) across Digit Loads (upper row), and for 4 Digits (middle row), time locked to the onset of the sentence-final adjective (0 msec). Raw effect waveforms (left column) are displayed along with contour maps of sample-level statistics (middle column) and raster plots of cluster-level statistics (right column). Clusters with an associated $p$-value below the specified threshold ($\alpha = 0.05$) are shown in blue shades; all other clusters (gray shades) were statistically not significant. ERP waveforms of midline electrodes (bottom row), time locked to the onset of the sentence-final adjective (0 msec). 2 Digit trials are shown in black, 4 Digit trials in red. Proportional quantifiers in thick lines, and Non-Proportional in thin.

epoch, particularly on left-hemispheric electrodes, possibly already around the N100-P200 components. Non-proportional quantifiers appear to be associated with a larger P200. Neither Quantifier Class shows a distinctive P300 component. Rather, the difference between the classes sustains throughout the epoch, with Proportional Quantifiers being more negative than Non-Proportional, particularly on temporal and centro-parietal electrodes of the left hemisphere (Figure 5).

Assessing these differences statistically, we found a broadly distributed, predominantly left-hemispheric, sustained negative effect (first-ranked negative cluster, NEG1: $T_{sum} = -5610.515$, $S = 1975$, $p = 0.017$) that lasts from approximately 260 to 500 msec. There were no effects of Digit Load, and no differences between 2 and 4 Digits within each quantifier class. Like for the sentence-final effects, we compared the different quantifier types within a class. None of the quantifier types (Aristotelian vs Numerical, Positive vs Negative Proportional) were significantly different overall or for either digit condition (2 or 4).

In summary, Quantifier Class is what is driving the sentence-internal effect. In particular, Proportional Quantifiers are associated with consistently more negative waveforms, particularly in the left hemisphere.

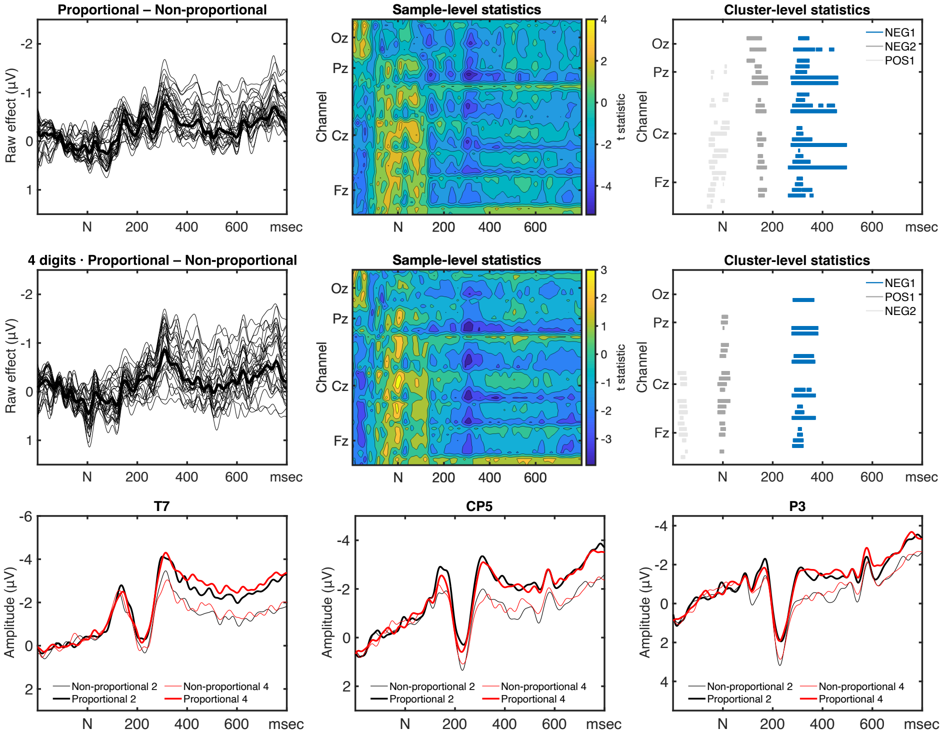


Figure 5: ERP effects of Quantifier Class (Proportional – Non-Proportional) across Digit Loads (upper row), and for 4 Digits (middle row), time locked to the onset of the sentence-internal noun (0 msec). Raw effect waveforms (left column) are displayed along with contour maps of sample-level statistics (middle column) and raster plots of cluster-level statistics (right column). Clusters with an associated $p$-value below the specified threshold ($\alpha = 0.05$) are shown in blue shades; all other clusters (gray shades) were statistically not significant. ERP waveforms of selected left-hemispheric electrodes (bottom row), time locked to the onset of the sentence-internal noun (0 msec). 2 Digit trials are shown in black, 4 Digit trials in red. Proportional quantifiers in thick lines, and Non-Proportional in thin.

There are some differences in the comparison between Quantifier Classes depending on Digit Load. While 4 Digit Proportional quantifiers are more negative than their 2 Digit counterparts, Non-Proportional quantifiers are more positive in the 4 Digit than in the 2 Digit case. As a result, the effect of Quantifier Class is larger for 4 Digits while the effect for 2 Digits does not reach significance.

### 3.2.3 Linear models of interactions between ERPs and individual WM scores

In order to ascertain whether the differences we found for the different Digit Loads and Truth Values were true interaction effects, we computed the individual mean cluster amplitude and $T_{sum}$ for each participant and constructed general linear models to assess significance.

At the noun, the linear model using mean amplitude in the first-ranked negative cluster did not reveal any significant effect (see Table 6). In particular, the interaction between Digit Load and Quantifier Class is not significant, and there were no significant main effects of WM measures on the ERPs, nor any significant interactions between WM measures and the two experimental manipulations. These results

| Condition | $\beta$ | SE | df | t | p |
|---|---|---|---|---|---|
| Intercept | -1.455 | 0.683 | 166.641 | -2.129 | 0.035 |
| Proportional | -0.711 | 0.791 | 135.000 | -0.899 | 0.370 |
| 4 Digits | 0.039 | 0.177 | 135.000 | 0.220 | 0.826 |
| Eriksen | 0.846 | 0.620 | 127.371 | 1.363 | 0.175 |
| Sternberg | -0.024 | 0.637 | 127.371 | -0.380 | 0.705 |
| Brown-Peterson | 1.094 | 0.616 | 127.371 | 1.777 | 0.078 |
| Quantifier Class × Digit Load | -0.079 | 0.250 | 135.000 | -0.315 | 0.754 |
| Quantifier Class × Eriksen | -0.099 | 0.272 | 135.000 | -0.364 | 0.716 |
| Quantifier Class × Sternberg | 0.129 | 0.279 | 135.000 | 0.461 | 0.646 |
| Quantifier Class × Brown-Peterson | -0.029 | 0.270 | 135.000 | -0.108 | 0.914 |
| Digit Load × Eriksen | -0.096 | 0.136 | 135.000 | -0.706 | 0.481 |
| Digit Load × Sternberg | -0.033 | 0.139 | 135.000 | -0.237 | 0.813 |
| Digit Load × Brown-Peterson | -0.131 | 0.135 | 135.000 | -0.970 | 0.334 |

Table 6: Linear mixed-effects model of mean amplitude in the first-ranked negative cluster at the noun for Proportional versus Non-Proportional Quantifiers.

| Condition | $\beta$ | SE | t | p |
|---|---|---|---|---|
| Intercept | -17.028 | 4.382 | -3.886 | $< 0.001$ |
| Eriksen | -7.360 | 4.762 | -1.545 | 0.129 |
| Sternberg | 1.746 | 4.888 | 0.357 | 0.723 |
| Brown-Peterson | -6.119 | 4.730 | -1.294 | 0.203 |

Table 7: Linear model of individual $T_{sum}$ in the first-ranked negative cluster at the noun for Proportional versus Non-Proportional Quantifiers Overall.

were replicated for the $T_{sum}$ analysis, where working memory scores had no significant impact on the difference between Proportional and Non-Proportional Quantifiers in either of the significant clusters (i.e., overall and 4 Digits). See Table 7 for the overall cluster, and *Supplementary material B*, section B.1, for the 4 Digit case.

At the sentence final adjective, the linear mixed-effects model of mean cluster amplitude in the first-ranked negative cluster for quantifier class and truth value revealed only significant main effects for Digit Load and Truth Value, and no significant interaction effects (see Table 8). In the regression on individual level $T_{sum}$s, only the intercept was significant, indicating that most of the variation is due to random individual differences. We report the result for the overall cluster in Table 9 and refer the reader to *Supplementary material B*, section B.2, for the same analysis of significant clusters by Digit Load and Truth Value.

# 4 Discussion

Overall, we found that memory load affects processing of Proportional and Non-Proportional Quantifiers differently. Both kinds of quantifiers exhibit a negative effect in the N200-N400 time-window for False *vs* True completions of the sentence, indicating that neural processes are sensitive to the truth value of the sentence at the final word. At the sentence-internal noun, we found a sustained negative effect of Proportional relative to Non-Proportional quantifiers, larger for 4 Digits than for 2.

Comparing these results with other reports in the literature, the sentence-final effects are consistent with those found in our previous experiment (Bremnes et al., 2022). The effect of Truth Value is earlier

| Condition | $\beta$ | SE | df | t | p |
|---|---|---|---|---|---|
| Intercept | 0.133 | 0.717 | 220.587 | 0.186 | 0.852 |
| Proportional | 0.443 | 0.815 | 316.000 | 0.543 | 0.587 |
| Digit Load | 0.450 | 0.173 | 316.000 | 2.589 | 0.010 |
| True | 2.026 | 0.348 | 316.000 | 5.829 | $< 0.0001$ |
| Eriksen | 0.992 | 0.779 | 220.587 | 1.274 | 0.204 |
| Sternberg | 0.290 | 0.799 | 220.587 | 0.363 | 0.717 |
| Brown-Peterson | 0.276 | 0.773 | 220.587 | 0.357 | 0.722 |
| Quantifier Class × Digit Load | -0.382 | 0.246 | 316.000 | -1.555 | 0.121 |
| Quantifier Class × Truth Value | -0.528 | 0.491 | 316.000 | -1.075 | 0.283 |
| Quantifier Class × Eriksen | -1.394 | 0.886 | 316.000 | -1.574 | 0.117 |
| Quantifier Class × Sternberg | -0.212 | 0.909 | 316.000 | -0.234 | 0.815 |
| Quantifier Class × Brown-Peterson | -0.970 | 0.880 | 316.000 | -1.103 | 0.271 |
| Digit Load × Eriksen | -0.096 | 0.189 | 316.000 | -0.510 | 0.611 |
| Digit Load × Sternberg | 0.029 | 0.194 | 316.000 | 0.150 | 0.881 |
| Digit Load × Brown-Peterson | 0.056 | 0.188 | 316.000 | 0.299 | 0.765 |
| Truth Value × Eriksen | -0.675 | 0.377 | 316.000 | -1.790 | 0.074 |
| Truth Value × Sternberg | -0.248 | 0.388 | 316.000 | -0.641 | 0.522 |
| Truth Value × Brown-Peterson | -0.026 | 0.375 | 316.000 | -0.069 | 0.945 |
| Quantifier Class × Digit Load × Eriksen | 0.369 | 0.267 | 316.000 | 1.381 | 0.168 |
| Quantifier Class × Digit Load × Sternberg | 0.045 | 0.274 | 316.000 | 0.165 | 0.869 |
| Quantifier Class × Digit Load × Brown-Peterson | 0.208 | 0.265 | 316.000 | 0.785 | 0.433 |
| Quantifier Class × Truth Value × Eriksen | 0.084 | 0.534 | 316.000 | 0.158 | 0.875 |
| Quantifier Class × Truth Value × Sternberg | -0.070 | 0.548 | 316.000 | -0.128 | 0.898 |
| Quantifier Class × Truth Value × Brown-Peterson | 0.253 | 0.530 | 316.000 | 0.477 | 0.634 |

Table 8: Linear mixed-effects model of mean amplitude in the first-ranked negative cluster at the adjective for Proportional versus Non-Proportional Quantifiers.

| Condition | $\beta$ | SE | t | p |
|---|---|---|---|---|
| Intercept | -25.088 | 5.319 | -4.717 | $< 0.0001$ |
| Eriksen | -5.724 | 5.780 | -0.990 | 0.327 |
| Sternberg | -4.130 | 5.932 | -1.172 | 0.490 |
| Brown-Peterson | -6.728 | 5.741 | -1.172 | 0.248 |

Table 9: Linear model of individual $T_{sum}$ in the first-ranked negative cluster at the adjective for Proportional versus Non-Proportional Quantifiers Overall.

than a traditional N400 (Augurzky et al., 2017; Vissers et al., 2008; Knoeferle et al., 2011). This effect can be followed by a positivity for more complex stimuli or tasks (Augurzky et al., 2017, 2019, 2020a,b), and we find indications of that in contrasts involving Proportional quantifiers. Early onset N400-like effects have been observed in contexts where semantic expectancy is very high (Van Petten et al., 1999), such as in the context of a picture (Vissers et al., 2008), but such early negativities have also been argued to reflect a mismatch between an active representation of the picture and the representation of the incoming sentence, manifesting as an N2b (D'Arcy et al., 2000; Wassenaar and Hagoort, 2007). Which of these interpretations turn out to be correct is inconsequential to our main argument, as both of them entail the completion of a verification procedure.

The sentence-internal effects described here are different from those we found in the previous study (Bremnes et al., 2022) and from those observed in earlier research on quantifier verification (Augurzky et al., 2020a; De Santo et al., 2019; Politzer-Ahles et al., 2013). These studies found positivities for proportional quantifiers, negative polarity, and semantic violations, respectively, while here we observed a negativity in the 250-500 msec time-window at the noun. Politzer-Ahles et al. (2013) did find a sustained negativity for pragmatic violations on quantifiers, but their effect was different both in terms of latency (500-1000 msec post-stimulus) and distribution (posterior) than our own negativity. The effect of Proportional quantifiers is more akin to the SANs observed for recomputation and ambiguity in discourse models (Baggio et al., 2008; Müller et al., 1997; Münte et al., 1998; van Berkum et al., 1999, 2003) or the LANs observed for long-distance dependencies (Fiebach et al., 2001; King and Kutas, 1995; Kluender and Kutas, 1995; Vos et al., 2001). Of particular note is the fact that such negativities have been reported to be modulated by working memory load (Vos et al., 2001).

Since our behavioral results are partially in line with earlier work (Szymanik and Zajenkowski, 2011; Zajenkowski and Szymanik, 2013; Zajenkowski et al., 2014), in that task performance is correlated with working memory scores, one might expect performance on the measures of executive function to correlate with the ERPs (Fiebach et al., 2002; Vos et al., 2001). However, no significant correlation was found. It is worth noting that the behavioral correlations are statistically weaker than those observed previously, and the Eriksen task did not correlate at all, contrary to previously reported effects (Zajenkowski and Szymanik, 2013; Zajenkowski et al., 2014).

## 4.1 Embedding the automata theory in the psychology of verification

It is important to distinguish the effects predicted by the automata theory from those that fall outside its purview. In particular, the modulation of the sentence-final effect is more likely to reflect a decision process based on the expectation that the sentence is true of the given picture, and not the unfolding of the verification procedure as such (Bremnes et al., 2022). Any interpretation of these effects can therefore only be inferred from the previous literature. By contrast, the negativity at the sentence-internal noun is plausibly related to the verification of the sentence, given that every other linguistic property was identical at this position in the sentence. The fact that we observed differences between the two Quantifier Classes at this position, and that these differences are more marked in the highest digit load condition, provides evidence that the observed on-line differences between Proportional and Non-proportional quantifiers are indeed related to memory resources. However, the direction of the interactions and the precise memory systems underlying them are not predicted by the theory, and interpretation thus remains speculative.

Bearing that in mind, the procedure that best explains our results is one in which participants build a model verifying the sentence on-line (Baggio, 2018; Clark and Chase, 1972, 1974; Clark, 1976; Johnson-Laird, 1983; Just and Carpenter, 1971; Just, 1974; Zwaan and Radvansky, 1998; van Lambalgen and Hamm, 2005), or proceed on the basis of the expectation that the picture provides a model for the sentence, i.e., that the sentence is true of the picture. One possibility here is that the brain entertains two models – one model of the picture, and one of the sentence – that it expects will conform to one another. The sentence model is being updated with each incoming word, and previous studies have shown that the picture model constrains the sentence model and gives very high semantic expectancy for the upcoming words (Augurzky et al. 2017; Knoeferle et al. 2014; Kuperberg 2016; Zwaan 2015; for evidence of the converse relation, see Coco et al. 2017). The incompatibility of the final word with this model of the sentence – i.e., the sentence matching the picture – is what is causing the N400-like activity observed for the False versus True comparison. This is true irrespective of whether this negativity is a true N400 or whether it reflects perceptual mismatch (Knoeferle et al., 2011; Vissers et al., 2008), as both alternatives presuppose the construction of a model for the sentence. It is therefore likely that these sentence-final effects reflect recomputation of the sentence model and/or decision-making processes (Augurzky et al., 2017; Knoeferle et al., 2014). The differences between Quantifier Classes at this point – compared to Non-Proportional, Proportional Quantifiers have a smaller N400, followed by a positivity for the False versus True comparison – do suggest that the entire process of verification, from determining the truth value to making a judgement, is affected by the complexity of the computational problem, as these effects are comparable to the effects of other kinds of complexity (Augurzky et al. 2017, 2020a; Politzer-Ahles et al. 2013; see also Nieuwland 2016; Urbach and Kutas 2010). However, the automata theory does not predict these differences, but only differences in determining the truth value. Importantly, in order to make a sentence model that is true of the sentence, one needs to know what completion of the sentence would make it true, which is equivalent to verifying the sentence. We therefore expect that the differences in the verification procedure predicted by the automata theory should occur prior to the effect of Truth Value. If participants are building a model of the sentence as the sentence unfolds, and this model is completed by the final word, as evidenced by the sentence-final Truth Value effect, then the difference between Quantifier Classes observed at the noun is plausibly an effect of differences in the verification procedure. The fact that these differences are modulated by Digit Load can therefore be taken as evidence that the verification procedure is modulated by memory.

Still, there are a couple of objections to such a view that are worth considering. It has been argued that while quantifiers are interpreted incrementally, their semantic representations are underspecified in such a way as to allow the final interpretation to occur significantly later, in particular in contexts where task demands are high, like in our case (Urbach and Kutas 2010; Urbach et al. 2015; see also Arcara et al. 2019). One conceivable alternative is therefore that the verification procedure is some kind of counting or estimation algorithm that returns numerosities, and that the actual verification happens only after adjective onset, where the participants are comparing the estimated numerosities of, e.g., all circles and all red circles. This would be an alternative explanation of the differences between quantifier classes at the adjective: instead of being downstream consequences of verification, they are direct verification effects. However, this does not change the complexity claims we set out to test: unbounded counting, which would be required by any quantifier without a specified numerical value, is not doable with an FSA (Hopcroft and Ullman, 1979).

More problematically, this alternative account leaves the effect of truth value unexplained. One could argue that there is an inherent cost to processing false, as opposed to true, sentences (Just and Carpenter, 1971; Clark and Chase, 1972, 1974), but that presupposes knowing the truth value. Since knowing the truth value is equivalent to having verified the sentence, the most likely explanation is that a verification processes has already been completed at the adjective, i.e., the participants predict the sentences to be a true description of the picture. The interpretation we are advocating provides an explanation of sentence-final effects in terms of violation of predictions. But if participants are not building a model, one should, in the absence of an alternative account of the differences, expect symmetry between true and false sentences, since the only difference between them is their truth and falsity relative to the model.

The burden is therefore on an alternative account to explain the observed asymmetry.

## 4.2 The implementation of the memory component

The fact that the sentence internal effect is different than the one observed previously warrants an explanation. As mentioned, the polarity of the effect is dependent on the orientation of the dipole generator, but the effect in the present study is different in both distribution and latency as well. This suggest that different memory components are involved depending on the task. For example, in the absence of the digit matching task, systems of recollection memory might suffice to perform the task, thus yielding an LPC-like effect (Rugg and Curran, 2007). By contrast, in the presence of the digit matching task, additional systems of working memory and executive function are recruited, resulting in ERP signatures traditionally associated with working memory in sentence processing, such as the SAN (Baggio et al., 2008; Müller et al., 1997; Münte et al., 1998; van Berkum et al., 1999, 2003) or sustained LAN (Fiebach et al., 2001, 2002; Vos et al., 2001). This could also explain the differences between Quantifier Classes by Digit Load, since the different nature of the kinds of verification algorithms (requiring or not requiring memory) potentially alters the task of verifying the sentence substantially enough to cause different memory systems to be recruited. On the basis of the results presented here, it is not possible to decide which memory systems (recollection memory, working memory) are engaged by verification of the different Quantifier Classes. Speculating, one possibility is that the negative effect of working memory effectively cancels the positive effect of recollection memory, i.e., that the negativity obscures a later positivity. Another possibility is that given a certain task complexity, the entire task is performed using a different memory system.

The data do not allow us to reverse inference which memory components are involved, but only give us new hypotheses to test. An important caveat for interpreting the present results is that while we observe an effect of Quantifier Class, the effect is different from the effects that have been observed previously. Whether this is the result of different memory systems being recruited, and if so, what causes different cognitive resources to be deployed in different tasks, remains an open question. Subsequent experiments should therefore be designed to answer these unresolved issues. A negative finding is that we could not correlate the ERPs to the working memory measures, as predicted by the theory. Future studies should further probe these correlations, possibly with other measures of working memory capacity, such as reading or digit span. The low variation, at least for some of the working memory tasks, does suggest that either (1) the tests are not valid because they are either too easy or too hard, so that the variation in the sample cannot be detected, or (2) the sample is too homogeneous. It might be that case that the population our sample comes from – i.e., university students – might not have enough spread in working memory capacity, and future research should aim at including a more diverse sample to explore whether the amplitude differences increase proportionally to the spread in the population. On the other hand, if the working memory battery we used was not appropriate, it might be possible to find correlations using more sensitive measures.

# 5 Conclusion

We have shown that the algorithmic complexity of a minimal verification algorithm is associated with different electrophysiological patterns, thus providing further evidence that psycholinguistics ought to be informed by results from theoretical computer science. One major limitation of the previous study (Bremnes et al., 2022) was that the relation to memory had to be inferred from the theory, and could not be demonstrated experimentally. The findings presented herein, however, suggest that the formal constraints applicable to abstract machines are not only applicable to but are of the same nature as the constraints on algorithms of human sentence processing.

It has been suggested that computational complexity analyses constitute an intermediate level between the computational and the algorithmic level (Isaac et al., 2014). These analyses should be able to assess whether posited computational problems are plausibly computable by the brain (van Rooij, 2008; van Rooij et al., 2019). Our results, here and in Bremnes et al. (2022), demonstrate that the minimal complexity of an algorithm delineates a lower bound on the algorithms used by the brain, regardless of their precise implementation. If, as our results indicate, the nature of the computational resources, e.g., a memory requirement, can be inferred from the formal theory, the space of possible algorithms used by the brain is considerably narrower. By observing that humans are constrained by computational resources derivable from formal theory and observable in the evoked potential, the Marrian perspective permits us to ignore computationally implausible hypotheses that would otherwise have to be tested. Consequently, the integration of formal and experimental results enables well-founded, plausible hypotheses that can likely reveal deep properties of the human capacity for language and cognition more generally (Bird, 2021; van Rooij and Baggio, 2020, 2021).

## Acknowledgements

JS has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERCGrant Agreement n. STG 716230 CoSaQ.

## References

Adam, K. C. S., Vogel, E. K., and Awh, E. (2020). Multivariate analysis reveals a generalizable human electrophysiological signature of working memory load. *Psychophysiology*, 57.

Amico, F., Ambrosini, E., Guillem, F., Mento, G., Power, D., Pergola, G., and Vallesi, A. (2015). The Virtual Tray of Objects Task as a novel method to electrophysiologically measure visuo-spatial recognition memory. *International Journal of Psychophysiology*, 98:477–489.

Arcara, G., Franzon, F., Gastaldon, S., Brotto, S., Semenza, C., Peressotti, F., and Zanini, C. (2019). One can be some but some cannot be one: ERP correlates of numerosity incongruence are different for singular and plural. *Cortex*, 116:104–121.

Augurzky, P., Bott, O., Sternefeld, W., and Ulrich, R. (2017). Are all the triangles blue? - ERP evidence for the incremental processing of German quantifier restriction. *Language and Cogntion*, 9:603–636.

Augurzky, P., Franke, M., and Ulrich, R. (2019). Gricean Expectations in Online Sentence Comprehension: An ERP Study on the Processing of Scalar Inferences. *Cognitive Science*, 43(8).

Augurzky, P., Hohaus, V., and Ulrich, R. (2020a). Context and Complexity in Incremental Sentence Interpretation: An ERP Study on Temporal Quantification. *Cognitive Science*, 44(11).

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

Axel, M. and Müller, N. G. (1996). Dissociations in the Processing of "What" and "Where" Information in Working Memory: An Event-Related Potential Analysis. *Journal of Cognitive Neuroscience*, 8:453–473.

Bach, E., Jelinek, E., Kratzer, A., and Partee, B. H., editors (1995). *Quantification in Natural Languages.* Kluwer Academic Publishers, Dordrecht.

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, 63:1–29.

Baggio, G. (2008). Processing temporal constraints: An erp study. *Language Learning*, 58(S1):35–55.

Baggio, G. (2018). *Meaning in the brain.* MIT Press, Cambridge, MA.

Baggio, G. (2020). Epistemic Transfer Between Linguistics and Neuroscience: Problems and Prospects. In Nefdt, R., Klippi, C., and Karstens, B., editors, *The Philosophy and Science of Language: Interdisciplinary Perspectives*, pages 275–308. Palgrave Macmillan, Cham.

Baggio, G., Stenning, K., and van Lambalgen, M. (2016). Semantics and cognition. *The Cambridge handbook of formal semantics*, pages 756–774.

Baggio, G., van Lambalgen, M., and Hagoort, P. (2008). Computing and recomputing discourse models: An ERP study. *Journal of Memory and Language*, 59:36–53.

Baggio, G., van Lambalgen, M., and Hagoort, P. (2015). Logic as marr's computational level: Four case studies. *Topics in Cognitive Science*, 7(2):287–298.

Bailey, K., Mlynarczyk, G., and West, R. (2016). Slow Wave Activity Related to Working Memory Maintenance in the N-Back Task. *Journal of Psychophysiology*, 30:141–154.

Barriga-Paulino, C. I., Rodríguez-Martínez, E. I., Rojas-Benjumea, M. Á., and Gómez, C. M. (2014). Slow wave maturation on a visual working memory task. *Brain and Cognition*, 88:43–54.

Barwise, J. and Cooper, R. (1981). Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, 4:159–219.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bird, A. (2021). Understanding the Replication Crisis as a Base Rate Fallacy. *The British Journal for the Philosophy of Science*, 74:965–993.

Bremnes, H. S., Szymanik, J., and Baggio, G. (2022). Computational complexity explains neural differences in quantifier verification. *Cognition*, 223:105013.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, 10:12–21.

Carcassi, F., Steinert-Threlkeld, S., and Szymanik, J. (2021). Monotone Quantifiers Emerge via Iterated Learning. *Cognitive Science*, 45.

Chemla, E., Dautriche, I., Buccola, B., and El Fagot, J. (2019). Constraints on the lexicons of human languages have cognitive roots present in baboons (Papio papio). *PNAS*, 116(30).

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2:113–124.

Clark, H. H. (1976). *Semantics and Comprehension*. Mouton, The Hague.

Clark, H. H. and Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, 3:472–517.

Clark, H. H. and Chase, W. G. (1974). Perceptual coding strategies in the formation and verification of descriptions. *Memory and Cognition*, 2:101–111.

Coco, M. I., Araujo, S., and Petersson, K. M. (2017). Disentangling stimulus plausibility and contextual congruency: Electro-physiological evidence for differential cognitive dynamics. *Neuropsychologia*, 96:150–163.

Coppock, E. (2019). Quantity Superlatives in Germanic, or "Life on the Fault Line Between Adjective and Determiner". *Journal of Germanic Linguistics*, 31:109–200.

D'Arcy, R. C. N., Connolly, J. F., and Crocker, S. F. (2000). Latency shifts in the N2b component track phonological deviations inspoken words. *Clinical Neurophysiology*, 111:40–44.

De Santo, A., Rawski, J., Yazdani, A. M., and Drury, J. E. (2019). Quantified Sentences as a Window into Prediction and Priming: An ERP Study. In Ronai, E., Stigliano, L., and Sun, Y., editors, *Proceedings of the Fifty-fourth Annual Meeting of the Chicago Linguistic Society*, pages 85–98, Chicago, IL. Chicago Linguistic Society.

Dehaene, S. (2011). *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press, Oxford, 2 edition.

Embick, D. and Poeppel, D. (2015). Towards a computational(ist) neurobiology of language: correlational, integrated and explanatory neurolinguistics. *Language, Cognition and Neuroscience*, 30(4):357–366.

Eriksen, B. A. and Eriksen, C. W. (1966). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception and Psychophysics*, 16:143–149.

Feigenson, L. (2005). A double-dissociation in infants' representations of object arrays. *Cognition*, 95:B37–B48.

Fiebach, C. J., Schlesewsky, M., and Friederici, A. D. (2001). Syntactic Working Memory and the Establishment of Filler-Gap Dependencies: Insights from ERPs and fMRI. *Journal of Psycholinguistic Research*, 30:321–338.

Fiebach, C. J., Schlesewsky, M., and Friederici, A. D. (2002). Separating syntactic memory costs and syntactic integration costs during parsing: the processing of German WH-questions. *Journal of Memory and Language*, 47:250–272.

Freunberger, D. and Nieuwland, M. S. (2016). Incremental comprehension of spoken quantifier sentences: Evidence from brain potentials. *Brain Research*, 1646:475–481.

Fukuda, K., Mance, I., and Vogel, E. K. (2015). $\alpha$ Power Modulation and Event-Related Slow Wave Provide Dissociable Correlates of Visual Working Memory. *Journal of Neuroscience*, 35:14009–14016.

García-Larrea, L. and Cézanne-Bert, G. (1998). P3, Positive slow wave and working memory load: a study on the functional correlates of slow wave activity. *Electroencephalography and Clinical Neurophysiology*, 108:260–273.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: most versus more than half. *Natural Language Semantics*, 17:63–98.

Harker, K. T. and Connolly, J. F. (2007). Assessment of visual working memory using event-related potentials. *Clinical Neurophysiology*, 118:2479–2488.

Hopcroft, J. E. and Ullman, J. D. (1979). *Introduction to automata theory, languages, and computation.* Addison-Wesley, Reading, Mass.

Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English.* PhD thesis, UCLA, Los Angeles, CA.

Hubbard, R. J., Rommers, J., Jacobs, C. L., and Federmeier, K. D. (2019). Downstream Behavioral and Electrophysiological Consequences of Word Prediction on Recognition Memory. *Frontiers in Human Neuroscience*, 13:291.

Hunt III, L., Politzer-Ahles, S., Gibson, L., Minai, U., and Fiorentino, R. (2013). Pragmatic inferences modulate N400 during sentence comprehension: Evidence from picture–sentence verification. *Neuroscience Letters*, 534:246–251.

Hunter, T. and Lidz, J. (2013). Conservativity and Learnability of Determiners. *Journal of Semantics*, 30:315–334.

Hunter, T., Lidz, J., Odic, D., and Wellwood, A. (2017). On how verification tasks are related to verification procedures: a reply to Kotek et al. *Natural Language Semantics*, 25:91–107.

Isaac, A. M. C., Szymanik, J., and Verbrugge, R. (2014). Logic and Complexity in Cognitive Science. In Baltag, A. and Smets, S., editors, *Johan van Benthem on Logic and Information Dynamics*, pages 787–824. Springer, Cham.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press, Cambridge.

Just, M. A. (1974). Comprehending quantified sentences: The relation between sentence-picture and semantic memory verification. *Cognitive Psychology*, 6:216–236.

Just, M. A. and Carpenter, P. A. (1971). Comprehension of Negation with Quantification. *Journal of Verbal Learning and Verbal Behavior*, 10:244–253.

Kanazawa, M. (2013). Monadic Quantifiers Recognized by Deterministic Pushdown Automata. In Aloni, M., Franke, M., and Roelofsen, F., editors, *Proceedings of the 19th Amsterdam Colloquium*, pages 139–146.

Keenan, E. and Stavi, J. (1986). A Semantic Characterization Of Natural Language Determiners. *Linguistics and Philosophy*, 9:253–326.

Keenan, E. L. and Paperno, D. (2017). Overview. In Paperno, D. and Keenan, E. L., editors, *Handbook of Quantifiers in Natural Language: Volume II*, pages 995–1004. Springer, Cham.

King, J. W. and Kutas, M. (1995). Who Did What and When? Using Word- and Clause-Level ERPs to Monitor Working Memory Usage in Reading. *Journal of Cognitive Neuroscience*, 7:376–395.

Kluender, J. W. and Kutas, M. (1995). Bridging the Gap: Evidence from ERPs on the Processing of Unbounded Dependencies. *Journal of Cognitive Neuroscience*, 5:196–214.

Knoeferle, P., Urbach, T. P., and Kutas, M. (2011). Comprehending how visual context influences incremental sentence processing: Insights from ERPs and picture-sentence verification. *Psychophysiology*, 48:495–506.

Knoeferle, P., Urbach, T. P., and Kutas, M. (2014). Different mechanisms for role relations versus verb–action congruence effects: Evidence from ERPs in picture–sentence verification. *Acta Psychologica*, 152:133–148.

Knowlton, T., Hunter, T., Odic, D., Wellwood, A., Halberda, J., Pietroski, P., and Lidz, J. (2021). Linguistic meanings as cognitive instructions. *Annals of the New York Academy of Sciences*.

Kounios, J. and Holcomb, P. (1992). Structure and Process in Semantic Memory: Evidence From Event-Related Brain Potentials and Reaction Times. *Journal of Experimental Psychology: General*, 121(4):459–479.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, 31:602–616.

Kusak, G., Grune, K., Hagendorf, H., and Metz, A.-M. (2000). Updating of working memory in a running memory task: an event-related potential study. *International Journal of Psychophysiology*, 39:51–65.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13):1–26.

Lefebvre, C., Vachon, F., Grimault, S., Thibault, J., Guimond, S., Peretz, I., Zatorre, R. J., and Jolicœur, P. (2013). Distinct electrophysiological indices of maintenance in auditory and visual short-term memory. *Neuropsychologia*, 51:2939–2952.

Lefebvre, C. D., Marchand, Y., Eskes, G. A., and Connolly, J. F. (2005). Assessment of working memory abilities using an event-related brain potential (ERP)-compatible digit span backward task. *Clinical Neurophysiology*, 116:1665–1680.

Levinson, S. C. (1983). *Pragmatics.* Cambridge University Press, Cambridge.

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature.* The MIT Press, Cambridge, MA.

Lewis, S. and Phillips, C. (2015). Aligning Grammatical Theories and Language Processing Models. *Journal of Psycholinguistic Research*, 44(1):27–46.

Lidz, J., Pietroski, P., Halberda, J., and Hunter, T. (2011). Interface transparency and the psychosemantics of most. *Natural Language Semantics*, 19:227–256.

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique.* The MIT Press, Cambridge, MA, 2 edition.

Luria, R., Balaban, H., Awh, E., and Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience & Biobehavioral Reviews*, 62:100–108.

Marchand, Y., D'Arcy, R. C. N., and Connolly, J. F. (2002). Linking neurophysiological and neuropsychological measures for aphasia assessment. *Clinical Neurophysiology*, 113:1715–1722.

Marchand, Y., Lefebvre, C. D., and Connolly, J. F. (2006). Correlating digit span performance and event-related potentials to assess working memory. *International Journal of Psychophysiology*, 62:280–289.

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164:177–190.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco.

Matthewson, L. (2001). Quantification and the Nature of Crosslinguistic Variation. *Natural Language Semantics*, 9:145–189.

McEvoy, L. K., Smith, M. E., and Gervins, A. (1998). Dynamic cortical networks of verbal and spatial working memory: Effects of memory load and task practice. *Cerebral Cortex*, 8:563–574.

Mostowski, M. (1998). Computational semantics for monadic quantifiers. *Journal of Applied Non-Classical Logics*, 8:107–121.

Müller, H. M., King, J. W., and Kutas, M. (1997). Event-related potentials elicited by spoken relative clauses. *Cognitive Brain Research*, 5:193–203.

Münte, T. F., Schiltz, K., and Kutas, M. (1998). When temporal terms belie conceptual order. *Nature*, 395:71–73.

Neath, I., Saint-Aubin, J., Bireta, T. J., Gabel, A. J., Hudson, C. G., and Surprenant, A. M. (2019). Short- and Long-Term Memory Tasks Predict Working Memory Performance, and Vice Versa. *Canadian Journal of Experimental Psychology*, 73:79–93.

Nieuwland, M. S. (2016). Quantification, Prediction, and the Online Impact of Sentence Truth-Value: Evidence From Event-Related Potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(2):316–334.

Noveck, I. A. and Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, 85(2):203–210.

Odic, D. and Starr, A. (2018). An Introduction to the Approximate Number System. *Child Development Perspectives*, 12:223–229.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011.

Partee, B. H. (2013). The Starring Role of Quantifiers in the History of Formal Semantics. In Punčochář, V. and Šarný, P., editors, *The Logica Yearbook 2012*, pages 113–136. College Publications.

Pelosi, L., Hayward, M., and Blumhardt, L. D. (1995). Is “memory-scanning” time in the Sternberg paradigm reflected in the latency of event-related potentials? *Electroencephalography and Clinical Neurophysiology*, 96:44–55.

Pelosi, L., Hayward, M., and Blumhardt, L. D. (1998). Which event-related potentials reflect memory processing in a digit-probe identification task? *Cognitive Brain Research*, 6:205–218.

Pelosi, L., Holly, M., Slade, T., Hayward, M., Barrett, G., and Blumhardt, L. D. (1992). Wave form variations in auditory event-related potentials evoked by a memory-scanning task and their relationship with tests of intellectual function. *Electroencephalography and Clinical Neurophysiology*, 84:344–352.

Peters, S. and Westerståhl, D. (2006). *Quantifiers in Language and Logic*. Oxford University Press, Oxford.

Peterson, L. and Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, 58:193–198.

Pietroski, P., Lidz, J., Hunter, T., and Halberda, J. (2009). The Meaning of 'Most': Semantics, Numerosity and Psychology. *Mind and Language*, 24:554–585.

Pietroski, P., Lidz, J., Hunter, T., Odic, D., and Halberda, J. (2011). Seeing what you mean, mostly. In Runner, J., editor, *Experiments at the Interfaces*, volume 37 of *Syntax and Semantics*, pages 181–217. Brill, Leiden.

Politzer-Ahles, S., Fiorentino, R., Jiang, X., and Zhou, X. (2013). Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification. *Brain Research*, 1490.

Quinlan, J. A., Neath, I., and Surprenant, A. M. (2015). Positional Uncertainty in the Brown-Peterson Paradigm. *Canadian Journal of Experimental Psychology*, 69:64–71.

Rai, M. K. and Harris, R. J. (2013). The Modified Brown-Peterson Task: A Tool to Directly Compare Children and Adult's Working Memory. *The Journal of Genetic Psychology*, 174:153–169.

Ratcliff, R., Sederberg, P. B., Smith, T. A., and Childers, R. (2016). A single trial analysis of EEG in recognition memory: Tracking the neural correlates of memory strength. *Neuropsychologia*, 93:128–141.

Rösler, F., Heil, M., and Röder, B. (1997). Slow negative brain potentials as reflections of specific modular resources of cognition. *Biological Psychology*, 45:109–141.

Ruchkin, D. S., Grafman, J., Cameron, K., and Berndt, R. S. (2003). Working memory retention systems: A state of activated long-term memory. *Behavioral and Brain Sciences*, 26:709–728.

Ruchkin, D. S., Johnson, R., Canoune, H., and Ritter, W. (1990). Short-term memory storage and retention: an event-related brain potential study. *Electroencephalography and Clinical Neurophysiology*, 76:419–439.

Ruchkin, D. S., Johnson, R., Grafman, J., Canoune, H., and Ritter, W. (1992). Distinctions and similarities among working memory processes: an event-related potential study. *Cognitive Brain Research*, 1:53–66.

Rugg, M. D. and Curran, T. (2007). Event-related potentials and recognition memory. *TRENDS in Cognitive Sciences*, 11:251–257.

Rugg, M. D., Mark, R. E., Walla, P., Schloerscheidt, A. M., Birch, C. S., and Allan, K. (1998). Dissociation of the neural correlates of implicit and explicit memory. *Nature*, 392:595—598.

Spychalska, M., Kontinen, J., Noveck, I., Reimer, L., and Werning, M. (2019). When numbers are not exact: Ambiguity and prediction in the processing of sentences with bare numerals. *Journal of Experimental Psychology: Learning Memory and Cognition*, 45(7):1177–1204.

Spychalska, M., Kontinen, J., and Werning, M. (2016). Investigating scalar implicatures in a truth-value judgement task: evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, 31:817–840.

Steinert-Threlkeld, S. and Szymanik, J. (2019). Learnability and semantic universals. *Semantics and Pragmatics*, 12.

Sternberg, S. (1966). High-Speed Scanning in Human Memory. *Science*, 153:652–654.

Szymanik, J. (2016). *Quantifiers and Cognition: Logical and Computational Perspectives.* Springer, Cham.

Szymanik, J. and Zajenkowski, M. (2010). Quantifiers and Working Memory. In Aloni, M., Bastiaanse, H., de Jager, T., van Ormondt, P., and Schulz, K., editors, *Amsterdam Colloquium 2009*, volume 25, pages 456–464, Berlin Heidelberg. Springer Verlag.

Szymanik, J. and Zajenkowski, M. (2011). Contribution of working memory in parity and proportional judgments. *Belgian Journal of Linguistics*, 25:176–194.

Talmina, N., Kochari, A., and Szymanik, J. (2017). Quantifiers and verification strategies: connecting the dots. In Cremers, A., van Gessel, T., and Roelofsen, F., editors, *Proceedings of the 21st Amsterdam Colloquium*, pages 465–473.

Tomaszewicz, B. (2011). Verification Strategies for Two Majority Quantifiers in Polish. In Reich, I., Horch, E., and Pauly, D., editors, *Proceedings of Sinn und Bedeutung 15*, pages 597–612.

Urbach, T. P., DeLong, K. A., and Kutas, M. (2015). Quantifiers are incrementally interpreted in context, more than less. *Journal of Memory and Language*, 83:79–96.

Urbach, T. P. and 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.

van Benthem, J. (1986). *Essays in Logical Semantics.* D. Reidel Publishing Company, Dordrecht.

van Berkum, J. J. A., Brown, C. M., and Hagoort, P. (1999). Early Referential Context Effects in Sentence Processing: Evidence from Event-Related Brain Potentials. *Journal of Memory and Language*, 41:147–182.

van Berkum, J. J. A., Brown, C. M., Hagoort, P., and Zwitserlood, P. (2003). Event-related brain potentials reflect discourse-referential ambiguity in spoken language comprehension. *Psychophysiology*, 40:235–248.

van de Pol, I., Steinert-Threlkeld, S., and Szymanik, J. (2019). Complexity and learnability in the explanation of semantic universals of quantifiers. In Goel, A. K., Seifert, C. M., and Freksa, C., editors, *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 3015–3021, Montreal, QB. Cognitive Science Society.

van Lambalgen, M. and Hamm, F. (2005). *The Proper Treatment of Events*. Blackwell, Malden.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., and Parks, M. (1999). Time Course of Word Identification and Semantic Integration in Spoken Language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25:394–417.

van Rooij, I. (2008). The Tractable Cognition Thesis. *Cognitive Science*, 32:939–984.

van Rooij, I. and Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychological Inquiry*, 31(4):321–325.

van Rooij, I. and Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science*.

van Rooij, I., Blokpoel, M., Kwisthout, J., and Wareham, T. (2019). *Cognition and Intractability: A Guide to Classical and Parameterized Complexity Analysis*. Cambridge Univeristy Press, Cambridge.

Vissers, C. T. W. M., Kolk, H. K. J., van de Meerendonk, N., and Chwilla, D. J. (2008). Monitoring in language perception: Evidence from ERPs in a picture–sentence matching task. *Neuropsychologia*, 46:967–982.

Vogel, E. K. and Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, 428:748–751.

Vos, S. H., Gunter, T. C., Kolk, H. H. J., and Mulder, G. (2001). Working memory constraints on syntactic processing: An electrophysiological investigation. *Psychophysiology*, 38:41–63.

Wassenaar, M. and Hagoort, P. (2007). Thematic role assignment in patients with Broca's aphasia: Sentence–picture matching electrified. *Neuropsychologia*, 45:716–740.

Yang, H., Laforge, G., Stojanoski, B., Nichols, E. S., McRae, K., and Köhler, S. (2019). Late positive complex in event-related potentials tracks memory signals when they are decision relevant. *Scientific Reports*, 9:9469.

Zajenkowski, M., Styła, R., and Szymanik, J. (2011). A computational approach to quantifiers as an explanation for some language impairments in schizophrenia. *Journal of Communication Disorders*, 44:595–600.

Zajenkowski, M. and Szymanik, J. (2013). MOST intelligent people are accurate and SOME fast people are intelligent. Intelligence, working memory, and semantic processing of quantifiers from a computational perspective. *Intelligence*, 41(5):456–466.

Zajenkowski, M., Szymanik, J., and Garraffa, M. (2014). Working Memory Mechanism in Proportional Quantifier Verification. *Journal of Psycholinguistic Research*, 43(6):839–853.

Zwaan, R. A. (2015). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, 23:1028–1034.

Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123:162–185.

# Paper 3

Neural Algorithms of Natural Language Quantification: A review of the experimental literature

This paper is awaiting publication and is not included in NTNU Open

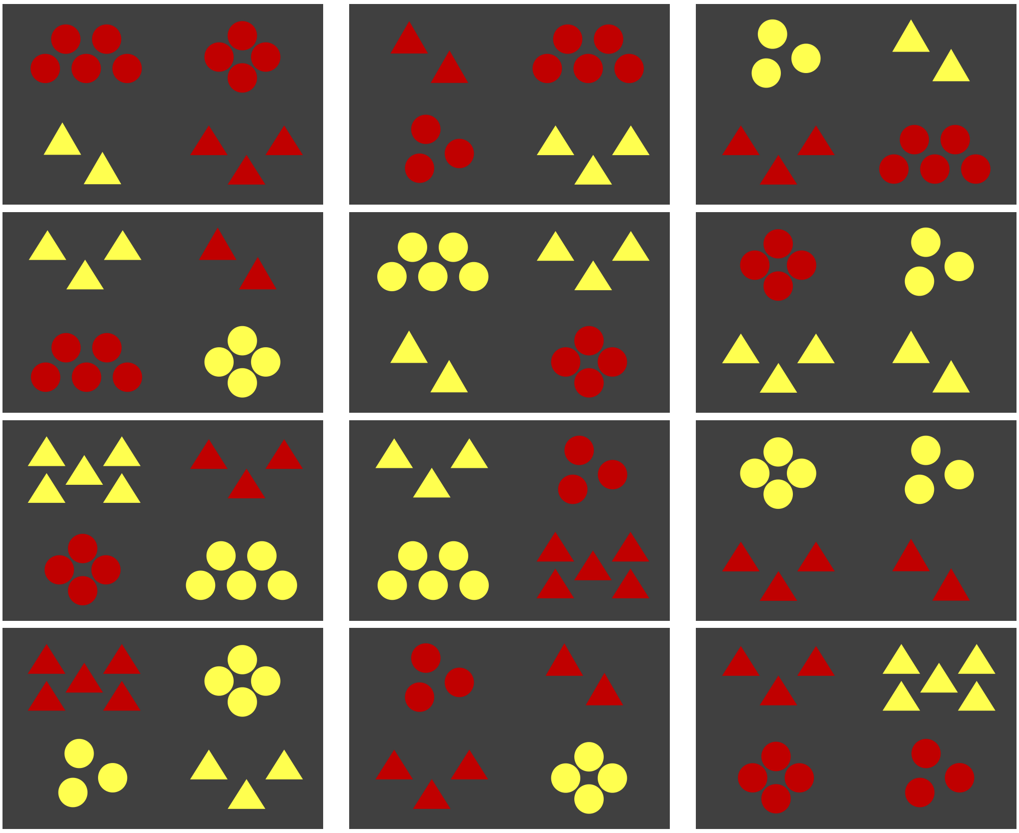
# Appendices

# Supplementary material

for *Computational complexity explains neural differences in quantifier verification*

# Contents

## 1 Images presented before the sentences



## 2 All experimental sentences

| | |
|---|---|
| **Proportional quantifiers** | |
| De fleste sirklene er gule | *Most circles are yellow* |
| De fleste sirklene er røde | *Most circles are red* |
| De fleste trekantene er gule | *Most triangles are yellow* |
| De fleste trekantene er røde | *Most triangles are red* |
| Færrest av sirklene er gule | *Fewest of the circles are yellow* |
| Færrest av sirklene er røde | *Fewest of the circles are red* |
| Færrest av trekantene er gule | *Fewest of the triangles are yellow* |
| Færrest av trekantene er røde | *Fewest of the triangles are red* |
| **Numerical quantifiers** | |
| Tre av sirklene er gule | *Three of the circles are yellow* |
| Tre av sirklene er røde | *Three of the circles are red* |
| Tre av trekantene er gule | *Three of the triangles are yellow* |
| Tre av trekantene er røde | *Three of the triangles are red* |
| Fem av sirklene er gule | *Five of the circles are yellow* |
| Fem av sirklene er røde | *Five of the circles are red* |
| Fem av trekantene er gule | *Five of the triangles are yellow* |
| Fem av trekantene er røde | *Five of the triangles are red* |
| **Aristotelian quantifiers** | |
| Alle sirklene er gule | *All the circles are yellow* |
| Alle sirklene er røde | *All the circles are red* |
| Alle trekantene er gule | *All the triangles are yellow* |
| Alle trekantene er røde | *All the triangles are red* |
| Ingen av sirklene er gule | *None of the circles are yellow* |
| Ingen av sirklene er røde | *None of the circles are red* |
| Ingen av trekantene er gule | *None of the triangles are yellow* |
| Ingen av trekantene er røde | *None of the triangles are red* |

## 3 Comprehension questions in Experiment 2

**Questions about the sentence**

| | |
|---|---|
| Er setninga en påstand om de fleste sirklene? | *Is the sentence a claim about most circles?* |
| Er setninga en påstand om færrest av sirklene? | *Is the sentence a claim about fewest of the circles?* |
| Er setninga en påstand om tre av sirklene? | *Is the sentence a claim about three of the circles?* |
| Er setninga en påstand om fem av sirklene? | *Is the sentence a claim about five of the circles?* |
| Er setninga en påstand om alle sirklene? | *Is the sentence a claim about all the circles?* |
| Er setninga en påstand om ingen av sirklene? | *Is the sentence a claim about none of the circles?* |

| | |
|---|---|
| Er setninga en påstand om de fleste trekantene? | *Is the sentence a claim about most triangles?* |
| Er setninga en påstand om færrest av trekantene? | *Is the sentence a claim about fewest of the triangles?* |
| Er setninga en påstand om tre av trekantene? | *Is the sentence a claim about three of the triangles?* |
| Er setninga en påstand om fem av trekantene? | *Is the sentence a claim about five of the triangles?* |
| Er setninga en påstand om alle trekantene? | *Is the sentence a claim about all the triangles?* |
| Er setninga en påstand om ingen av trekantene? | *Is the sentence a claim about none of the triangles?* |

| | |
|---|---|
| Er setninga en påstand om gule sirkler? | *Is the sentence a claim about yellow circles?* |
| Er setninga en påstand om gule trekanter? | *Is the sentence a claim about yellow triangles?* |
| Er setninga en påstand om røde sirkler? | *Is the sentence a claim about red circles?* |
| Er setninga en påstand om røde trekanter? | *Is the sentence a claim about red triangles?* |

| | |
|---|---|
| Er setninga en påstand om sirkler? | *Is the sentence a claim about circles?* |
| Er setninga en påstand om trekanter? | *Is the sentence a claim about triangles?* |

**Questions about the picture**

| | |
|---|---|
| Er det gule sirkler på bildet? | *Are there yellow circles in the picture?* |
| Er det gule trekanter på bildet? | *Are there yellow triangles in the picture?* |
| Er det røde sirkler på bildet? | *Are there red circles in the picture?* |
| Er det røde trekanter på bildet? | *Are there red triangles in the picture?* |

**Questions about both the picture and the sentence**

| | |
|---|---|
| Er det gule sirkler både på bildet og i setninga? | *Are there yellow circles both in the picture and in the sentence?* |
| Er det gule trekanter både på bildet og i setninga? | *Are there yellow triangles circles both in the picture and in the sentence?* |
| Er det røde sirkler både på bildet og i setninga? | *Are there red circles both in the picture and in the sentence?* |
| Er det røde trekanter både på bildet og i setninga? | *Are there red triangles circles both in the picture and in the sentence?* |

| | |
|---|---|
| Er det sirkler både på bildet og i setninga? | *Are there circles both in the picture and in the sentence?* |
| Er det trekanter både på bildet og i setninga? | *Are there triangles circles both in the picture and in the sentence?* |

# Supplementary material A

for *The interplay of computational complexity and memory load during quantifier verification*

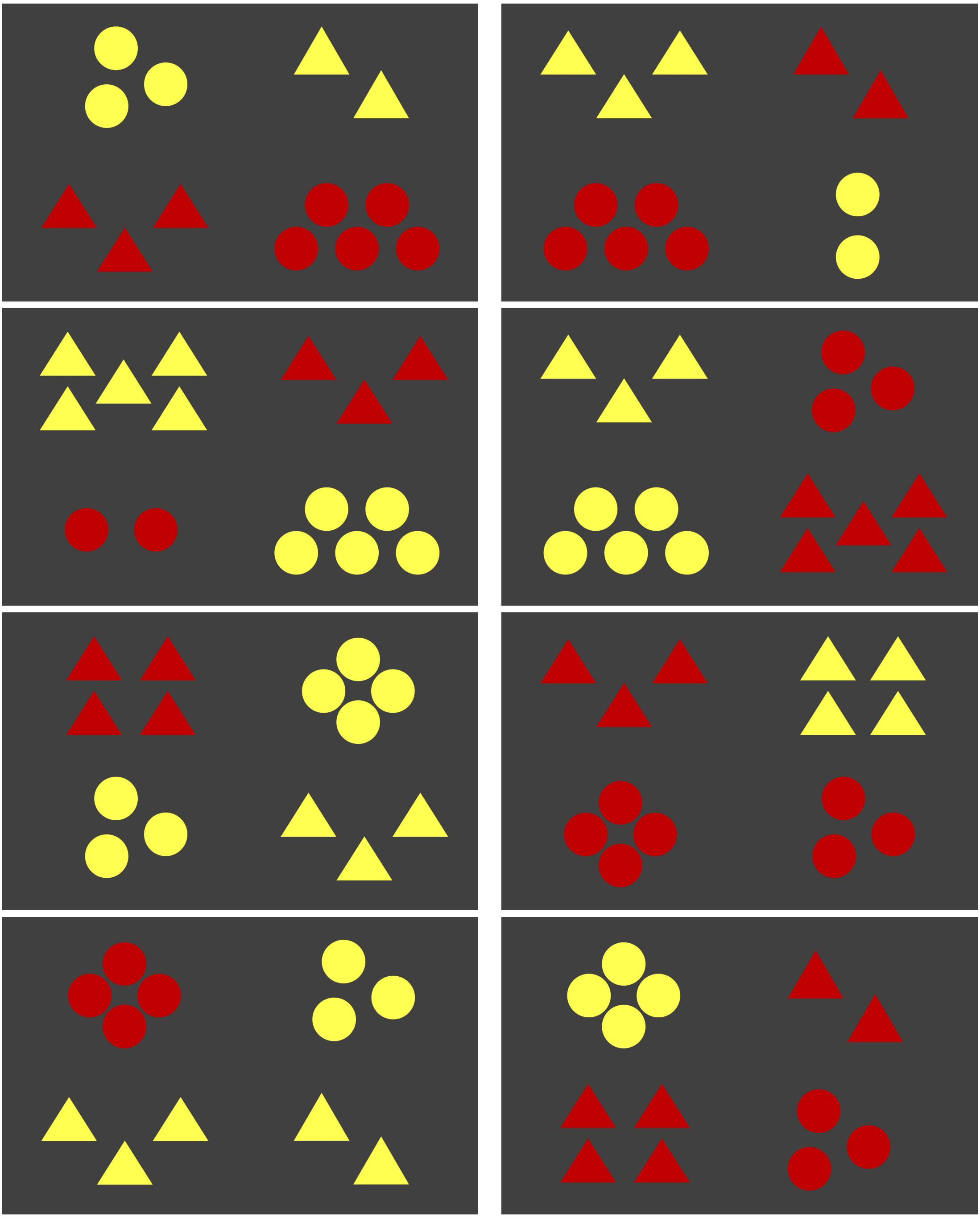
# Supplementary material A

*Memory Load Interacts with Computational Complexity in the Neural Signals of Quantifier Verification*

Heming Strømholt Bremnes, Jakub Szymanik, and Giosuè Baggio

## Contents

# I Images presented before the sentences



# II All experimental sentences

**Non-Proportional Quantifiers**

| | | |
|---|---|---|
| Samtlige av sirklene er røde | *All of the circles are red* | |
| Samtlige av sirklene er gule | *All of the circles are yellow* | |
| Samtlige av trekantene er røde | *All of the triangles are red* | |
| Samtlige av trekantene er gule | *All of the triangles are yellow* | |
| Ingen av sirklene er røde | *None of the circles are red* | |
| Ingen av sirklene er gule | *None of the circles are yellow* | |
| Ingen av trekantene er røde | *None of the triangles are red* | **Aristotelian** |
| Ingen av trekantene er gule | *None of the triangles are yellow* | |
| Enkelte av sirklene er røde | *Some of the circles are red* | |
| Enkelte av sirklene er gule | *Some of the circles are yellow* | |
| Enkelte av trekantene er røde | *Some of the triangles are red* | |
| Enkelte av trekantene er gule | *Some of the triangles are yellow* | |
| Tre av sirklene er røde | *Three of the circles are red* | |
| Tre av sirklene er gule | *Three of the circles are yellow* | |
| Tre av trekantene er røde | *Three of the triangles are red* | |
| Tre av trekantene er gule | *Three of the triangles are yellow* | |
| Fire av sirklene er røde | *Four of the circles are red* | |
| Fire av sirklene er gule | *Four of the circles are yellow* | |
| Fire av trekantene er røde | *Four of the triangles are red* | **Numerical** |
| Fire av trekantene er gule | *Four of the triangles are yellow* | |
| Fem av sirklene er røde | *Five of the circles are red* | |
| Fem av sirklene er gule | *Five of the circles are yellow* | |
| Fem av trekantene er røde | *Five of the triangles are red* | |
| Fem av trekantene er gule | *Five of the triangles are yellow* | |

**Proportional Quantifiers**

| | | |
|---|---|---|
| Flertallet av sirklene er røde | *The majority of the circles are red* | **Positive** |
| Flertallet av sirklene er gule | *The majority of the circles are yellow* | |
| Flertallet av trekantene er røde | *The majority of the triangles are red* | |
| Flertallet av trekantene er gule | *The majority of the triangles are yellow* | |
| Flest av sirklene er røde | *Most of the circles are red* | |
| Flest av sirklene er gule | *Most of the circles are yellow* | |
| Flest av trekantene er røde | *Most of the triangles are red* | |
| Flest av trekantene er gule | *Most of the triangles are yellow* | |
| Over halvparten av sirklene er røde | *More than half of the circles are red* | |
| Over halvparten av sirklene er gule | *More than half of the circles are yellow* | |
| Over halvparten av trekantene er røde | *More than half of the triangles are red* | |
| Over halvparten av trekantene er gule | *More than half of the triangles are yellow* | |
| Mindretallet av sirklene er røde | *The minority of the circles are red* | **Negative** |
| Mindretallet av sirklene er gule | *The minority of the circles are yellow* | |
| Mindretallet av trekantene er røde | *The minority of the triangles are red* | |
| Mindretallet av trekantene er gule | *The minority of the triangles are yellow* | |
| Færrest av sirklene er røde | *The fewest circles are red* | |
| Færrest av sirklene er gule | *The fewest circles are yellow* | |
| Færrest av trekantene er røde | *The fewest triangles are red* | |
| Færrest av trekantene er gule | *The fewest triangles are yellow* | |
| Under halvparten av sirklene er røde | *Less than half of the circles are red* | |
| Under halvparten av sirklene er gule | *Less than half of the circles are yellow* | |
| Under halvparten av trekantene er røde | *Less than half of the triangles are red* | |
| Under halvparten av trekantene er gule | *Less than half of the triangles are yellow* | |

# Supplementary material B

for *The interplay of computational complexity and memory load during quantifier verification*

# Supplementary material B

*The interplay of computational complexity and memory load during quantifier verification*
Heming Strømholt Bremnes, Jakub Szymanik, and Giosuè Baggio

## Contents

## A Inferential statistics on behavioral data

Mixed effects logistic and linear regression models for the accuracy and reaction time data, respectively, were constructed using the glmer function of the lme4 package (Bates et al., 2015) in R. The sentence verification task and the digit recollection task were modeled separately. Fixed effects were condition (Proportional/Non-Proportional for the linguistic task, and 2/4 Digits for digit recollection) and interaction (2/4 digits for sentence verification, and Proportional/Non-Proprotional for digit matching) and the measures of executive function. The measures of executive function were z-transformed overall accuracy ($z = \frac{x-m}{sd}$) for the Sternberg and Brown-Peterson tasks, and the z-transformed median reaction time difference between congruent and incongruent trials in the Eriksen flanker task. The models had random intercepts by participant and individual quantifier.

For verification task accuracy, $\beta$-estimates from the logistic regression (see table 1) revealed that participants were marginally more accurate with Proportional quantifiers. Furthermore, higher accuracy on the Brown-Peterson and Sternberg tasks was associated with higher accuracy, and participants were also more accurate with True sentences than with False. Lastly, accuracy was independent of digit load, meaning that there was no significant difference between storing 2 or 4 digits in memory.

In the linear regression on reaction time (see table 2), $\beta$-estimates revealed no effect of Quantifier Class, but participants were significantly faster with True sentences than with False. None of the working memory measures were associated with a significant increase or decrease in reaction time, thus suggesting that the variance is not related to individual differences in working memory capacity. We attempted to include random intercepts by randomization - i.e., presentation order - but they did not explain any of the variance, and were therefore omitted in the final model to avoid over-fitting.

| Condition | $\beta$ | SE | z | p |
|---|---|---|---|---|
| Intercept | 2.286 | 0.120 | 19.041 | $< .0001$ |
| Eriksen | -0.035 | 0.091 | -0.388 | 0.698 |
| Sternberg | 0.192 | 0.092 | 2.089 | 0.037 |
| Brown-Peterson | 0.252 | 0.092 | 2.743 | 0.006 |
| Proportional | 0.220 | 0.112 | 1.968 | 0.049 |
| 4 Digits | 0.070 | 0.077 | 0.910 | 0.363 |
| True | 0.321 | 0.077 | 4.146 | $< .001$ |

Table 1: Logistic regression on accuracy, Sentence verification task

Turning to the digit matching task, the logistic regression (see table 3) revealed that participants were significantly less accurate with 4 Digits, compared to 2. Higher accuracy on the Sternberg and Brown-Peterson tasks, as well as a larger difference median reaction time difference between congruent and incongruent trials in

| Condition | $\beta$ | SE | t | df | p |
|---|---|---|---|---|---|
| Intercept | 1565.653 | 98.837 | 7.419 | 15.841 | $< .0001$ |
| Eriksen | 48.828 | 74.141 | 43.995 | 0.659 | 0.513 |
| Sternberg | -59.798 | 76.092 | 43.995 | -0.786 | 0.436 |
| Brown-Peterson | -32.691 | 73.631 | 43.992 | -0.444 | 0.659 |
| Proportional | 218.999 | 99.746 | 2.000 | 2.196 | 0.159 |
| 4 Digits | -7.097 | 22.839 | 9126.665 | -0.311 | 0.756 |
| True | -61.407 | 22.679 | 9136.100 | -2.708 | 0.007 |

Table 2: Linear regression on response times, Sentence verification task

the Eriksen flanker task, parametrically increased accuracy, and trials were the digit pairs matched were also more likely to elicit a correct response in the digit task.

For reaction times, $\beta$-estimates from a linear regression revealed a significant difference between 2 and 4 Digits, such that participants were significantly slower with 4 Digits. They also responded faster in matching trials. Quantifier Class in the sentence verification task did not modulate response times in the Digit task, and, in contrast to accuracy, response times were not related to any working memory score.

| Condition | $\beta$ | SE | z | p |
|---|---|---|---|---|
| Intercept | 2.439 | 0.111 | 21.888 | $< .0001$ |
| Eriksen | 0.204 | 0.096 | 2.118 | 0.034 |
| Sternberg | 0.262 | 0.099 | 2.656 | 0.008 |
| Brown-Peterson | 0.274 | 0.097 | 2.815 | 0.005 |
| 4 Digits | -0.321 | 0.072 | -4.483 | $< .0001$ |
| Proportional | 0.055 | 0.078 | 0.710 | 0.478 |
| Match | 0.209 | 0.071 | 2.923 | 0.003 |

Table 3: Logistic regression on accuracy, Digit matching task

| Condition | $\beta$ | SE | t | df | p |
|---|---|---|---|---|---|
| Intercept | 1664.081 | 58.771 | 38.664 | 28.315 | $< .0001$ |
| Eriksen | 11.602 | 58.254 | 43.996 | 0.199 | 0.843 |
| Sternberg | -86.835 | 59.787 | 43.995 | -1.452 | 0.153 |
| Brown-Peterson | -5.153 | 57.853 | 43.994 | -0.089 | 0.929 |
| 4 Digits | 224.864 | 19.073 | 9143.687 | 11.790 | $< .0001$ |
| Proportional | -31.851 | 31.405 | 2.014 | -1.014 | 0.417 |
| Match | -285.887 | 19.122 | 8913.745 | -14.950 | $< .0001$ |

Table 4: Linear regression on response times, Digit matching task

# B Linear models of interactions between ERPs and individual WM scores

## B.1 Sentence-internal noun

| Condition | $\beta$ | SE | t | p |
|---|---|---|---|---|
| Intercept | -15.028 | 4.243 | -3.636 | $< 0.001$ |
| Eriksen | -9.702 | 4.611 | -2.104 | 0.041 |
| Sternberg | -0.010 | 4.732 | -0.021 | 0.983 |
| Brown-Peterson | -6.309 | 4.579 | -1.378 | 0.175 |

Table 5: Linear model of individual $T_{sum}$ in the first-ranked negative cluster at the noun for 4 Digit Proportional versus 4 Digit Non-Proportional Quantifiers.

## B.2 Sentence-final adjective

| Condition | $\beta$ | SE | t | p |
|---|---|---|---|---|
| Intercept | -60.250 | 15.940 | -3.780 | $< 0.001$ |
| Eriksen | 15.010 | 17.32 | 0.866 | 0.391 |
| Sternberg | 15.110 | 17.78 | 0.850 | 0.400 |
| Brown-Peterson | 11.86 | 17.200 | 0.690 | 0.494 |

Table 6: Linear model of individual $T_{sum}$ in the first-ranked negative cluster at the adjective for Proportional versus Non-Proportional Quantifiers in True trials.

| Condition | $\beta$ | SE | t | p |
|---|---|---|---|---|
| Intercept | -80.755 | 19.961 | -4.046 | $< 0.001$ |
| Eriksen | 13.280 | 21.692 | 0.612 | 0.544 |
| Sternberg | -12.368 | 22.263 | -0.556 | 0.581 |
| Brown-Peterson | -2.566 | 21.543 | -0.119 | 0.906 |

Table 7: Linear model of individual $T_{sum}$ in the first-ranked negative cluster at the adjective for Proportional versus Non-Proportional Quantifiers for 4 Digits.

| Condition | $\beta$ | SE | t | p |
|---|---|---|---|---|
| Intercept | -41.271 | 11.167 | -3.696 | $< 0.001$ |
| Eriksen | 10.616 | 12.135 | 0.875 | 0.386 |
| Sternberg | 18.523 | 12.454 | 1.487 | 0.144 |
| Brown-Peterson | -7.036 | 12.052 | -0-584 | 0.562 |

Table 8: Linear model of individual $T_{sum}$ in the first-ranked negative cluster at the adjective for Proportional versus Non-Proportional Quantifiers in 2 Digit True trials.

| Condition | $\beta$ | SE | t | p |
|---|---|---|---|---|
| Intercept | -108.818 | 30.924 | -3.519 | 0.001 |
| Eriksen | 10.148 | 33.605 | 0.302 | 0.764 |
| Sternberg | 0.573 | 34.489 | 0.017 | 0.987 |
| Brown-Peterson | 37.469 | 33.374 | 1.123 | 0.268 |

Table 9: Linear model of individual $T_{sum}$ in the first-ranked negative cluster at the adjective for Proportional versus Non-Proportional Quantifiers in 4 Digit True trials.

| Condition | $\beta$ | SE | t | p |
|---|---|---|---|---|
| Intercept | -76.530 | 21.740 | -3.520 | 0.001 |
| Eriksen | 3.629 | 23.625 | 0.154 | 0.879 |
| Sternberg | -4.198 | 24.247 | -0.173 | 0.863 |
| Brown-Peterson | -4.911 | 23.463 | -0.209 | 0.835 |

Table 10: Linear model of individual $T_{sum}$ in the first-ranked negative cluster at the adjective for Proportional versus Non-Proportional Quantifiers in 4 Digit False trials.

ISBN 978-82-326-7110-6  (printed ver.)
ISBN 978-82-326-7109-0 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology