

The interplay of computational complexity and memory load during quantifier verification

Heming Strømholth Bremnes, Jakub Szymanik & Giosuè Baggio

To cite this article: Heming Strømholth Bremnes, Jakub Szymanik & Giosuè Baggio (2023): The interplay of computational complexity and memory load during quantifier verification, *Language, Cognition and Neuroscience*, DOI: [10.1080/23273798.2023.2236253](https://doi.org/10.1080/23273798.2023.2236253)

To link to this article: <https://doi.org/10.1080/23273798.2023.2236253>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 07 Aug 2023.



[Submit your article to this journal](#)




[View related articles](#)



[View Crossmark data](#)

The interplay of computational complexity and memory load during quantifier verification

Heming Strømholth Bremnes ^a, Jakub Szymanik ^b and Giosuè Baggio ^a

^aLanguage Acquisition and Language Processing Lab, Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway; ^bCenter for Mind/Brain Sciences, Department of Information Engineering and Computer Science, University of Trento, 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. Additionally, an interaction between Digit-Load and Quantifier-Class was observed at the sentence-final word. Our results suggest that constraints on cognitive resources deployed during human sentence processing and verification are of the same nature as formal constraints on abstract machines.

ARTICLE HISTORY

Received 3 June 2022
Accepted 30 June 2023

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 & Paperno, 2017; Matthewson, 2001). Pioneering work (Barwise & Cooper, 1981; Keenan & 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 & Lidz, 2013; Steinert-Threlkeld & Szymanik, 2019; van de Pol et al., 2023). 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 step-wise 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., 2016, 2015; Embick & Poeppel, 2015; Lewis & Phillips, 2015), specifying the properties of the algorithms that underlie cognitive computation is essential. It may therefore seem puzzling that algorithmic aspects of on-line human semantic processing hitherto

CONTACT Heming Strømholth Bremnes  heming.s.bremnes@ntnu.no

 Supplemental data for this article can be accessed online at <http://dx.doi.org/10.1080/23273798.2023.2236253>.

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

have not received sufficient attention (Baggio, 2018, 2020). One reason for this might be the fact that meanings are notoriously hard to formalise 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 formalised in *generalized quantifier theory* as relations between the cardinalities of sets (Barwise & Cooper, 1981; Peters & Westerståhl, 2006). This approach has made quantifiers a linchpin in the development of formal semantics (Partee, 2013) and has enabled the construction of verification algorithms for quantifiers, to be discussed in more detail in Section 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 generalised quantifier theory (Barwise & Cooper, 1981; Keenan & Stavi, 1986) as (conservative and extensional) relations between cardinalities of sets allows determiner meanings to be modeled as sets of strings of binary recognised 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 & 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 probably require a memory component where an algorithm can store information about red and non-red circles, and therefore require the additional computational resources of a *pushdown automaton* (PDA) for their verification. 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 & Nieuwland, 2016; Kounios & Holcomb, 1992; Nieuwland, 2016; Noveck & Posada, 2003; Urbach et al., 2015; Urbach & Kutas, 2010), and

several have used a picture-sentence verification task to study the processing of quantified sentences (Augurzky et al., 2017, 2019; Augurzky, Hohaus, et al., 2020; Augurzky, Schlotterbeck, et al., 2020; Hunt III et al., 2013; Politzer-Ahles et al., 2013; Spychalska et al., 2019, 2016). 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 for our current purpose, 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 Hubbard et al., 2019; Ratcliff et al., 2016; Rugg et al., 1998; Yang et al., 2019). This effect is observed when participants are recalling contextual details of a stimulus (Rugg & Curran, 2007) and 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., 1995, 1998, 1992), or other digit span tasks (C. D. Lefebvre et al., 2005; Marchand et al., 2006), and have been argued to index

retrieval of information from short-term memory (García-Larrea & 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 & Müller, 1996; Rösler et al., 1997; Ruchkin et al., 1990, 1992; Vogel & 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 & Kutas, 1995; Kluender & Kutas, 1995; Vos et al., 2001). Sustained left-anterior negativities seem more likely candidates for 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 & Connolly, 2007; C. 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 vs 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 & Connolly, 2007), increased SAN amplitudes have been associated with greater auditory short-term memory capacity (C. 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). These results demonstrate 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 Section 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 logically possible 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 factors 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. The hypothesis is that if the difference is related to memory, then we will observe a difference in the evoked potential as a function of the memory manipulation. If we do not observe a difference in the evoked potential, then, by *modus tollens*, the difference is not related to memory.

On the basis of previously observed behavioural effects (Zajenkowski et al., 2011; Zajenkowski & Szymanski, 2013; Zajenkowski et al., 2014), we expect individual differences in the preliminary tasks to correlate with ERP signals. However, the direction of this correlation is not predicted, as working memory capacity and ERP effects 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 attempts to make explicit the 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, one can 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 & 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 does not.

2. Methods

2.1. Design

The study used a $2 \times 2 \times 2$ 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 as 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 memorise 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 analysed 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. 455 334).

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 & 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 & 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 s, 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 s and flattens out after that, so that there is only a small accuracy difference between, e.g. 12 and 18 s (Rai & Harris, 2013).

As mentioned in Section 2.1 above, the main tasks were to memorise 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 condition. Previous studies (Szymanik & Zajenkowski, 2010, 2011) found that, with 4 and 6 digits, digit recall was poor at 6 digits. In contrast, performance in 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 memorised 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 colour of these shapes were varied pseudorandomly. The grid design with a 2×2 potential shape by colour 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 colour, the other was always in different colours. 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 sentences, in which a certain colour was predicated of a certain quantity of shapes (e.g. “Flest av sirkene 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 colour 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 and 96 trials with 2/4 digits. There were 48 trials in each cell in the $2 \times 2 \times 2$ design. This number is standard in ERP

Table 1. The experimental sentences were constructed by combining every element of one column with every element of the other columns, resulting in $12 \times 2 \times 1 \times 2 = 48$ 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.

Quantifier Class	Quantifier	Shape	Copula	Color
Aristotelian	Samtlige av			
	Ingen av	sirkene <i>the</i> circles		røde <i>red</i>
Numerical	Enkelte av			
	Tre av			
	Fire av			
Positive	Fem av		er	
	Flesteparten av		are	
Proportional	Flest av			
	Over halvparten av	trekantene		gule <i>yellow</i>
Negative	Minsteparten av	<i>the</i>		
	Færrest av	triangler		
	Under halvparten av			

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 determined 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 (for each sentence with a given quantifier, there are 16 possible True/False combinations when considering Digit Load and the combination of nouns and adjectives), 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 memorised 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 colour, 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 colour. 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 colour 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 colour 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 randomised within each block. To counterbalance sentence types within a block, we also constructed 2 randomised 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 Section 2.5 below. After 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 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 every 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 s, preceded and followed by 500 msec of blank screen and a 500 msec fixation cross. Next, the picture was presented for

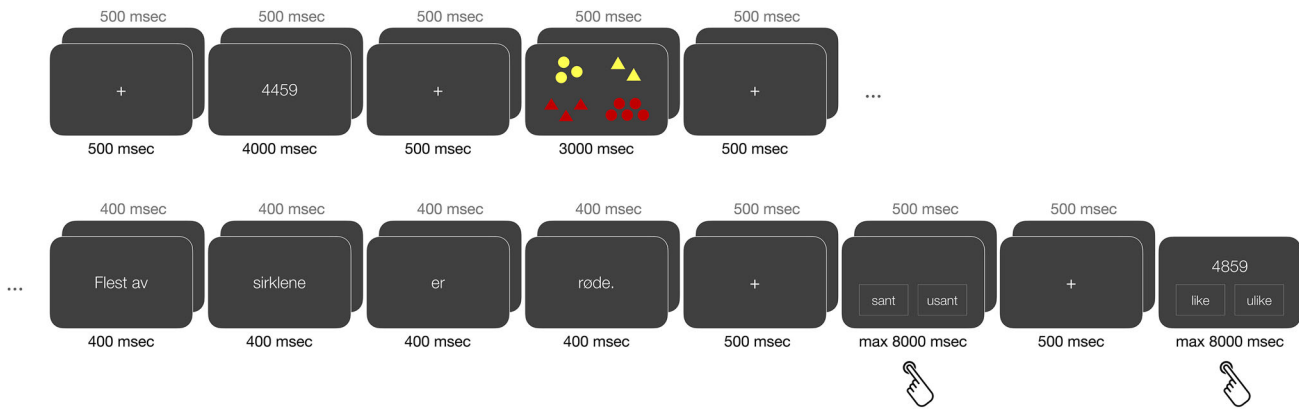


Figure 1. Structure of a single trial.

3 s, 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 blank screen in between. 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 but 8 s 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 h, including the preliminary tests (20–25 min), EEG setup (30–40 min), and the main experiment with breaks (1:10–1:30 h).

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 1000Hz 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.

2.6. Data analysis

Accuracy and reaction time data were collected for both the sentence verification and the digit recollection tasks, also in order to compare our results with those of previous behavioural studies. Accuracy was used to ensure that participants were performing the task correctly. Note that reaction times here 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 analysed using FieldTrip (Oostenveld et al., 2011). 1000 msec epochs, with a 200 msec pre-stimulus baseline, were extracted at the noun and at the sentence-final adjective. Trials with voltage values exceeding $\pm 150\mu\text{V}$ 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 analysed using non-parametric cluster-based statistics (Maris & Oostenveld, 2007), using the default alpha thresholds (.05) at both the sample and cluster levels. To assess ERP differences between two conditions, each sample (channel-time pair) was compared by means of a t -test. Adjacent samples passing a test were added to form a cluster, and their t -values were summed (T_{sum}). 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 randomly partitioning it into two subsets of equal size. 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 adopted two approaches. Firstly, we generated ERPs of the differences by subtracting the Non-Proportional ERP from the Proportional ERP for each digit condition, i.e. 2 Digit Proportional – 2 Digit Non-Proportional and 4 Digit Proportional – 4 Digit Non-Proportional. Subsequently we assessed the significance of the interaction by comparing the 2 Digit difference to the 4 Digit difference by means of the same non-parametric cluster-based algorithm described above.¹ This procedure was conducted both at the sentence-final adjective and at the sentence-internal noun. Secondly, in order to test the association between the pretest scores and the interactions, 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 we constructed models with these as the dependent variable, instead of mean 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 behavioural 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

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	
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 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.

	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

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 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.

Table 4. Descriptive statistics for 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.

	M	SD
Eriksen	62.250	32.356
Sternberg	0.866	0.072
Brown-Peterson	0.383	0.182

Table 5. Correlation matrix of behavioural and working memory measures, where DAcc = Digit Accuracy, DRT = Digit RT, QPAcc = Proportional quantifier accuracy, QNPAcc = Non-Proportional quantifier accuracy, BP = Brown-Peterson task.

	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

Notes: 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.

3.2. ERP results

3.2.1. Sentence-final effects: adjective

We began by analysing 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 ERPs for True and False sentences begin to reconverge around 450 msec. This waveform difference is also reflected in the statistics (Figure 2): we see a broadly distributed negative effect of False vs True (first-ranked negative cluster, NEG1: $T_{sum} = -16,685.102$, $S = 3629$, $p = 0.001$). 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. This effect is confirmed by statistical analysis (Figure 3). We found a positive cluster (first-ranked positive cluster, POS1: $T_{sum} = 2356.829$, $S = 929$,

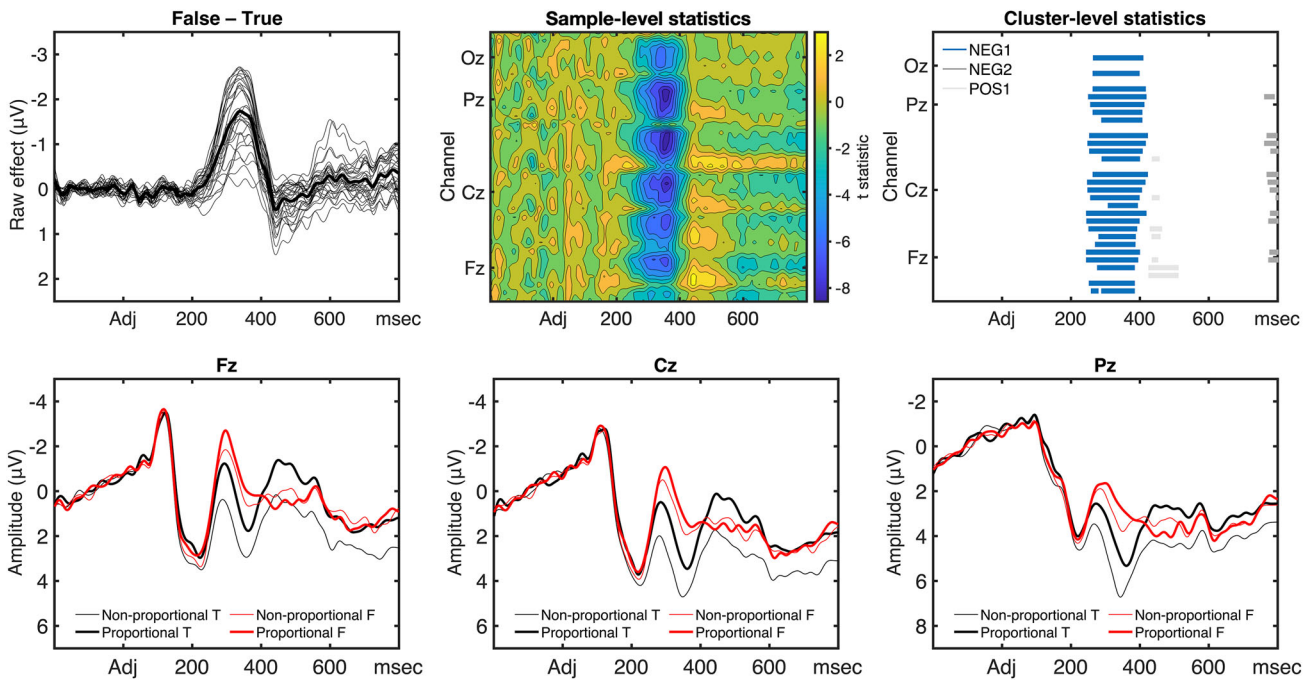


Figure 2. ERP effects of truth value (False–True) across quantifier classes (upper row), time locked to the onset of the sentence-final adjective (0 msec). Raw effect waveforms (upper left) are displayed along with contour maps of sample-level statistics (upper middle) and raster plots of cluster-level 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 at midline electrodes (bottom row), time locked to the onset of the sentence-final adjective (0 msec).

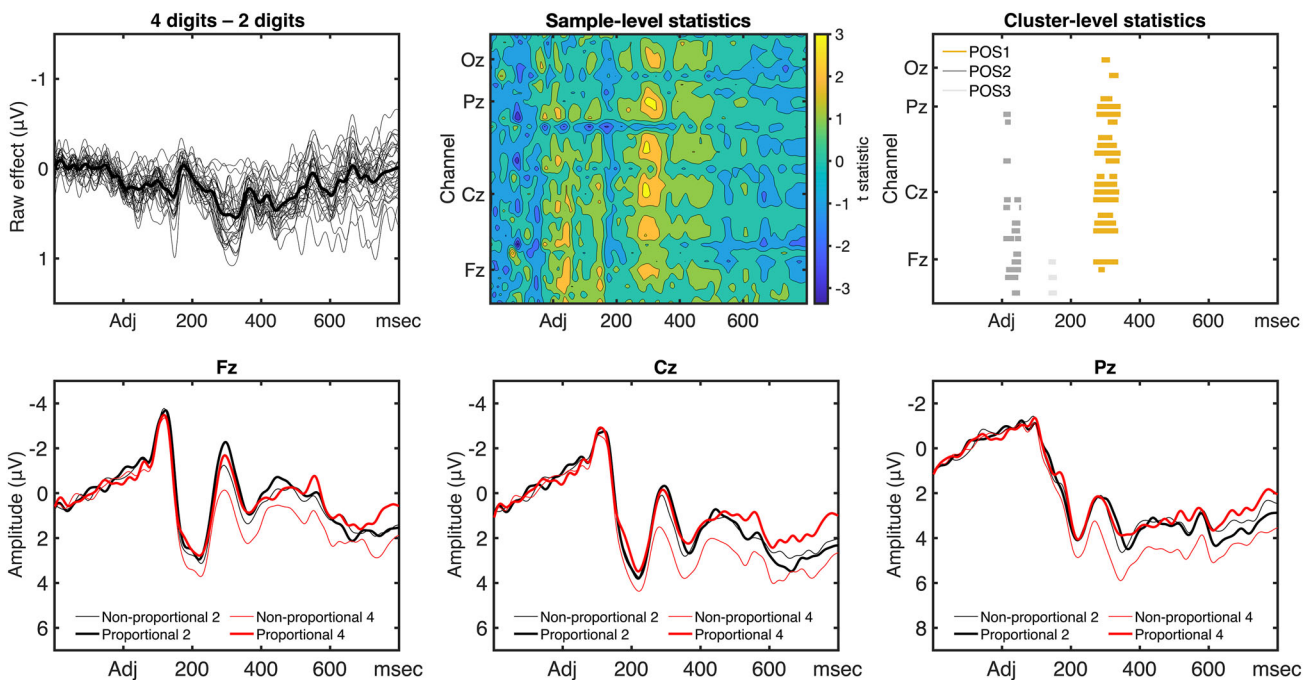


Figure 3. ERP effects of Digit Load (2 Digits–4 Digits) across quantifier classes (upper row), time locked to the onset of the sentence-final adjective (0 msec). Raw effect waveforms (upper left) are displayed along with contour maps of sample-level statistics (upper middle) and raster plots of cluster-level 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 at midline electrodes (bottom row), time locked to the onset of the sentence-final adjective (0 msec).

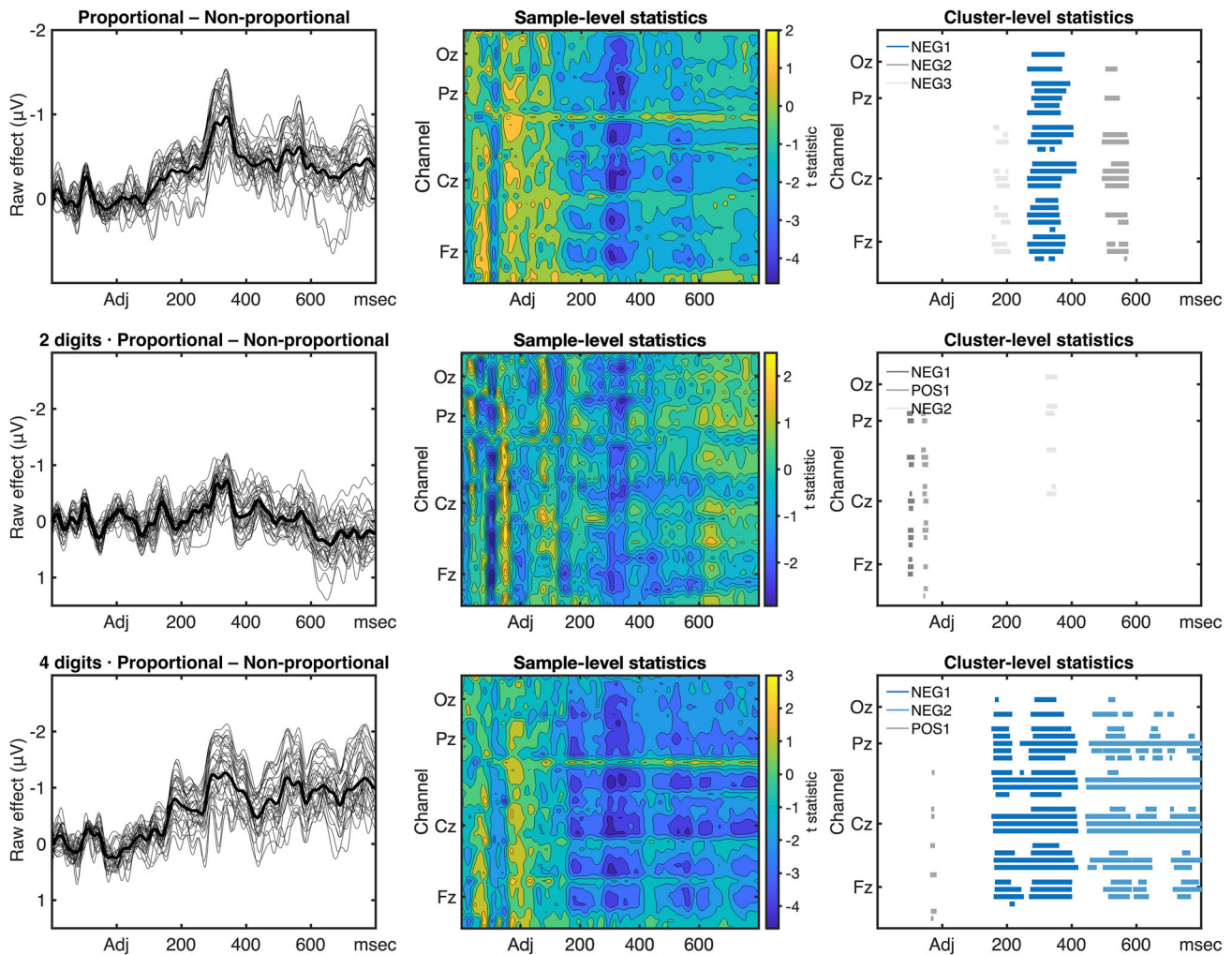


Figure 4. ERP effects of Quantifier Class (Proportional–Non-Proportional) across Digit Loads (upper row), and for 2 Digits (middle row) and 4 Digits (bottom row), time locked to the onset of the sentence-final adjective (0 msec). Raw effect waveforms (left column) are displayed along with contour maps 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.

$p = 0.049$) with a central, but more posterior distribution around 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: $T_{sum} = -12,537.21$, $S = 4294$, $p = 0.002$; NEG2: $T_{sum} = -10,599.67$, $S = 3960$, $p = 0.004$), which were not found in the 2 Digit condition (no significant

clusters). We also compared positive and negative Proportional Quantifiers to make sure 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, ERP effects are 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 in the analysis.

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 epoch, particularly over left-hemispheric electrodes, possibly already around the N100–P200 components. Nouns following non-proportional quantifiers appear to be associated with a larger P200, but neither Quantifier Class shows distinctive P200, P300 or N400 effects. Rather, the difference between the classes sustains throughout the epoch, with nouns after Proportional Quantifiers being more negative than after Non-Proportional quantifiers, 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 statistically significant 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. There are some differences in the comparison between Quantifier Classes depending on Digit Load: The effect of Quantifier Class is larger for 4 Digits than for 2 (NEG1: $T_{sum} = -2383.680$, $S = 884$, $p = 0.046$), and the effect for 2 Digits does not reach significance (NEG1: $T_{sum} = -1703.766$, $S = 697$, $p = 0.071$). The reason that the 4 Digit case is statistically weaker than the overall effect—in terms of T_{sum} , size, and p -value—despite being larger than the 2 Digit case, is presumably reduced statistical power resulting from only having half as many trials as in the overall comparison.

3.2.3. Interaction effects

In order to ascertain whether the differences we found for the different Digit Loads and Truth Values were true interaction effects, we constructed difference waves for Proportional and Non-Proportional at each digit condition, and compared the 2 Digit difference to the 4 Digit difference. At the noun, the cluster algorithm did not reveal a significant interaction (no significant positive or negative clusters), indicating that the difference at the noun is primarily modulated by Quantifier Class. By contrast, the same comparison at the adjective revealed one significant and one borderline significant positive cluster (POS1: $T_{sum} = 6528.903$, $S = 2667$, $p = 0.013$; POS2: $T_{sum} = 1627.001$, $S = 643$, $p = 0.081$). Both clusters are centrally distributed, and the largest cluster lasts from 450 ms after adjective onset to the end of the epoch, and the smaller is found between 160 and 230 ms. This means that there is a significantly larger negative difference between Proportional and Non-Proportional quantifiers in the higher Digit Load condition after the effect of Truth Value (see Figure 4).

3.2.4. Linear models of interactions between ERPs and individual WM scores

To explore the potential relationship between the ERPs and their interactions with the pretest scores, 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

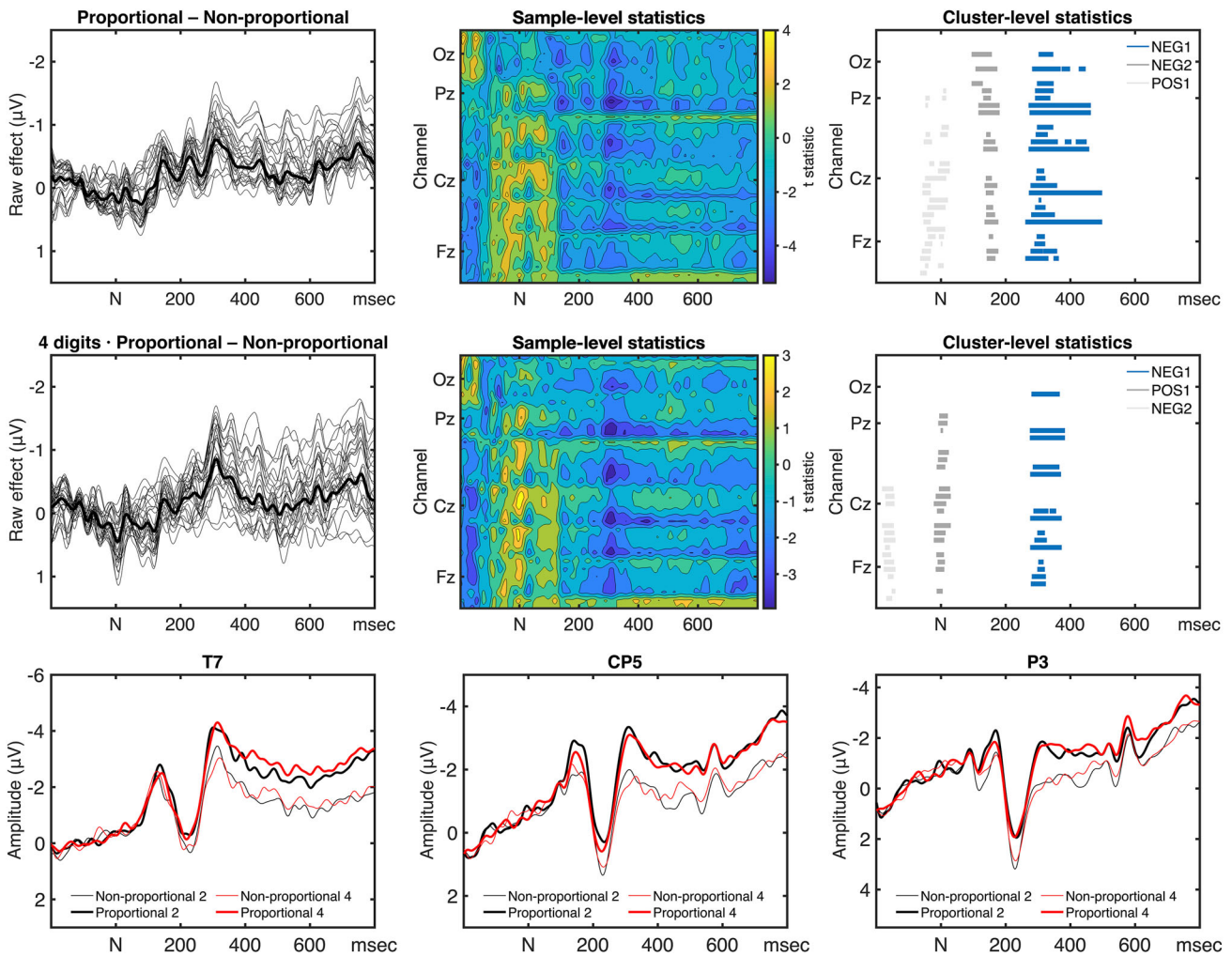


Figure 5. ERP effects of Quantifier Class (Proportional–Non-Proportional) across Digit Loads (upper row), and for 4 Digits (middle row), time locked to the onset of the sentence-internal noun (0 msec). Raw effect waveforms (left column) are displayed along with contour maps of sample-level statistics (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 at selected left-hemispheric electrodes (bottom row), time locked to the onset of the sentence-internal noun (0 msec).

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. The latter results 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

Table 6. Linear mixed-effects model of mean amplitude in the first-ranked negative cluster at the noun for Proportional vs Non-Proportional Quantifiers.

Condition	β	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 \times Digit Load	-0.079	0.250	135.000	-0.315	0.754
Quantifier Class \times Eriksen	-0.099	0.272	135.000	-0.364	0.716
Quantifier Class \times Sternberg	0.129	0.279	135.000	0.461	0.646
Quantifier Class \times Brown-Peterson	-0.029	0.270	135.000	-0.108	0.914
Digit Load \times Eriksen	-0.096	0.136	135.000	-0.706	0.481
Digit Load \times Sternberg	-0.033	0.139	135.000	-0.237	0.813
Digit Load \times Brown-Peterson	-0.131	0.135	135.000	-0.970	0.334

Table 7. Linear model of individual T_{sum} in the first-ranked negative cluster at the noun for Proportional vs Non-Proportional Quantifiers Overall.

Condition	β	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

effects (see Table 8). In the regression on individual level T_{sum} , 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 classes 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 shortly after presentation of the final word. Moreover, after the Truth Value effect, there is a larger negative difference between Proportional and Non-Proportional for 4 Digits than for 2. At the sentence-internal noun, we found a sustained negative effect of Proportional relative to Non-Proportional quantifiers, but no statistically reliable interaction was found with Digit Load.

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 than a traditional N400 (Augurzyk et al., 2017; Knoeferle et al., 2011; Vissers et al., 2008). 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 & 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 truth value effect can be followed by a positivity for more complex stimuli or tasks (Augurzyk et al., 2017, 2019; Augurzyk, Hohaus, et al., 2020; Augurzyk, Schlotterbeck, et al., 2020), and we find indications of that in contrasts involving Proportional quantifiers. However, when truth value is factored out, the interaction analysis reveals that, at the sentence-final adjective, there is a negative difference between Proportional and Non-proportional quantifiers for 4 Digits which is not found for their 2 Digit counterparts. Such negative shifts have previously been associated with recomputation of discourse models or revision of a discourse-level inference (Baggio et al., 2010, 2008; Pijnacker et al., 2011; Politzer-Ahles et al., 2013).

The sentence-internal effects described here are different from those we found in the previous study

Table 8. Linear mixed-effects model of mean amplitude in the first-ranked negative cluster at the adjective for Proportional vs Non-Proportional Quantifiers.

Condition	β	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 \times Digit Load	-0.382	0.246	316.000	-1.555	0.121
Quantifier Class \times Truth Value	-0.528	0.491	316.000	-1.075	0.283
Quantifier Class \times Eriksen	-1.394	0.886	316.000	-1.574	0.117
Quantifier Class \times Sternberg	-0.212	0.909	316.000	-0.234	0.815
Quantifier Class \times Brown-Peterson	-0.970	0.880	316.000	-1.103	0.271
Digit Load \times Eriksen	-0.096	0.189	316.000	-0.510	0.611
Digit Load \times Sternberg	0.029	0.194	316.000	0.150	0.881
Digit Load \times Brown-Peterson	0.056	0.188	316.000	0.299	0.765
Truth Value \times Eriksen	-0.675	0.377	316.000	-1.790	0.074
Truth Value \times Sternberg	-0.248	0.388	316.000	-0.641	0.522
Truth Value \times Brown-Peterson	-0.026	0.375	316.000	-0.069	0.945
Quantifier Class \times Digit Load \times Eriksen	0.369	0.267	316.000	1.381	0.168
Quantifier Class \times Digit Load \times Sternberg	0.045	0.274	316.000	0.165	0.869
Quantifier Class \times Digit Load \times Brown-Peterson	0.208	0.265	316.000	0.785	0.433
Quantifier Class \times Truth Value \times Eriksen	0.084	0.534	316.000	0.158	0.875
Quantifier Class \times Truth Value \times Sternberg	-0.070	0.548	316.000	-0.128	0.898
Quantifier Class \times Truth Value \times Brown-Peterson	0.253	0.530	316.000	0.477	0.634

Table 9. Linear model of individual T_{sum} in the first-ranked negative cluster at the adjective for Proportional vs Non-Proportional Quantifiers Overall.

Condition	β	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

(Bremnes et al., 2022) and from those observed in earlier research on quantifier verification (Augurzky, Hohaus, et al., 2020; De Santo et al., 2019; Politzer-Ahles et al., 2013). These studies found positivities for proportional quantifiers, for negative polarity expressions, and for semantic violations, 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 & Kutas, 1995; Kluender & 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 behavioural results are partially in line with earlier work (Szymanik & Zajenkowski, 2011; Zajenkowski & 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 behavioural correlations are statistically weaker than those observed previously, and the Eriksen task did not correlate at all, contrary to previously reported effects (Zajenkowski & Szymanik, 2013; Zajenkowski et al., 2014).

4.1. Embedding the automata theory in the psychology of verification

In our previous study, we argued that an effect of truth value on ERPs indicates that participants have implicitly already determined whether the sentence is true or false (Bremnes et al., 2022). The interaction with memory we observed here is predominantly after the truth value effect. Since these later effects are modulated by

memory load, they are also potential candidates for neural instantiations of the abstract automata memory, thereby seemingly casting doubt on our original interpretation of the time-course of the verification process.

One possibility is therefore that participants wait until the proposition is completed and only subsequently initiate the verification procedure. On this view, the verification process starts when there is some evidence for either truth or falsity, and ends as soon as one of them is chosen. This view of verification contrasts with the automata view, where the entire computation of a semantic automaton is the verification process: i.e. the processing of all the objects denoted by the noun phrase, for each object deciding whether it has a property or not, and making a decision once all the objects in the domain have been classified. Nevertheless, 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 et al., 2015; Urbach & Kutas, 2010; see also Arcara et al., 2019). Another conceivable alternative, that is more compatible with the automata view, is therefore that the procedure initiated at the noun 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. Note that 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 & Ullman, 1979).

However, this account leaves the effect of truth value unexplained, since the verification procedure is only instantiated at the adjective and seemingly subsumes the truth value effect, at least in the 4 digit case. One could argue that there is an inherent cost to processing false, as opposed to true, sentences (Chang, 1986; Clark & Chase, 1972, 1974; Just & Carpenter, 1971), but that presupposes knowing the truth value. Since knowing the truth value entails having verified the sentence, a more plausible explanation is that a verification processes has already been completed at the adjective, i.e. the participants predict the sentence to be a true description of the picture. The interpretation we have previously adopted provides an explanation of sentence-final effects in terms of violation of predictions.

But if participants are not building a model in which the sentence is expected to come out true of the picture, 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.

Bearing that in mind, we still maintain that the procedure that best explains our results, and that is most compatible with other findings, is one in which participants build 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), 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 vs 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 possible that these sentence-final effects reflect computation 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—i.e. the interaction between Quantifier Class and Digit Load—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. This interpretation also explains the interaction effect at the adjective, since the negative shift has been associated with ambiguity in, revision of and difficulty of integration into discourse models (Baggio et al., 2010, 2008; Pijnacker et al., 2011; van Berkum et al., 1999, 2003). The same is true of the modulation of the truth value effects by Quantifier Class—compared to Non-Proportional, Proportional Quantifiers have a smaller N400, followed by a positivity for the False vs True

comparison—as these effects are comparable to the effects of other kinds of complexity (Augurzky et al., 2017; Augurzky, Hohaus, et al., 2020; Politzer-Ahles et al., 2013; see also Nieuwland, 2016; Urbach & Kutas, 2010). However, the automata theory does not predict these differences, but only differences in determining the truth value. Importantly, in order to build a sentence model that is true of the sentence, one needs to know what completion of the sentence would make it true, which involves 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, understood as per the automata theory.

The fact that these differences did not significantly interact with Digit Load is problematic for the automata view. One interpretation is that memory load builds incrementally as sentence processing commences, and that it is only when the additional resources for making a decision are recruited that the interaction effect of memory load is visible in the evoked potential. This is in line with the prediction from Section 1.3 that if there are larger differences in the 4 Digit condition, then this could be explained by proportional quantifiers straining participants' cognitive capacity. Such an interpretation is supported by previous findings where effects that have a similar spatiotemporal distribution to our interaction effect have been related to the increased effort of integrating more complex information into the wider context (Baggio et al., 2010, 2008; Pijnacker et al., 2011), which in this case is the picture. Alternatively, it is not uncommon for sustained effects to increase over time (see, e.g. Hagoort, 2003, and references therein), so that a difference originating at the sentence-internal noun might only reach significance at the sentence-final adjective, which is the next position of measurement.

Needless to say, there are at present no sentence processing models that neatly explain all the effects we observe. At this point, it is important to distinguish the effects predicted by the automata theory from those that fall outside its purview. The theory predicts there to be a qualitative difference between proportional and non-proportional quantifiers, which is confirmed by our ERP results. Specifically, the difference should be related to memory, and we therefore hypothesised that manipulating memory load should lead to an interaction between Digit Load and Quantifier class. This

hypothesis was also corroborated. The other effects are not within the predictive scope of the automata theory, and any interpretation of these effects can consequently only be inferred from the previous literature. In particular, the time-course of verification, the direction of the interactions and the precise memory systems underlying them are not predicted by the theory, and interpretation thus remains speculative.

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 suggests 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 & 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 infer 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. There are also some marginal length differences between quantifiers (2 out of 6 proportional quantifiers were 3 rather than 2 words), which may have impacted processing at the noun following the quantifier. However, as mentioned in the methods section, we were not able to look at ERPs at the level of individual quantifiers. A negative finding is that we could not correlate the ERPs to the working memory or executive function 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.

5. Conclusion

We have shown that the algorithmic complexity of a minimal verification algorithm is associated with different electrophysiological patterns, thus providing a strong argument that the psychology and neuroscience of language and reasoning 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 could not be demonstrated experimentally and had to be inferred from the theory. The findings presented herein, however, suggest that the formal constraints on abstract machines are not only also 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 & Baggio, 2020, 2021).

Note

1. We thank an anonymous reviewer for this suggestion.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

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.

Data availability statement

Scripts and data for this paper are available open access at DataverseNO (doi: [10.18710/LODKF](https://doi.org/10.18710/LODKF)) (Bremnes et al., 2023).

ORCID

Heming Strømholth Bremnes  <http://orcid.org/0000-0002-6390-5611>

Jakub Szymanik  <http://orcid.org/0000-0002-6145-6322>

Giosuè Baggio  <http://orcid.org/0000-0001-5086-0365>

References

- Adam, K. C. S., Vogel, E. K., & Awh, E. (2020). Multivariate analysis reveals a generalizable human electrophysiological signature of working memory load. *Psychophysiology*, *57*. <https://doi.org/10.1111/psyp.v57.12>
- Amico, F., Ambrosini, E., Guillem, F., Mento, G., Power, D., Pergola, G., & 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. <https://doi.org/10.1016/j.ijpsycho.2015.10.006>
- Arcara, G., Franzon, F., Gastaldon, S., Brotto, S., Semenza, C., Peressotti, F., & 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. <https://doi.org/10.1016/j.cortex.2018.10.022>
- Augurzky, P., Bott, O., Sternefeld, W., & Ulrich, R. (2017). Are all the triangles blue? – ERP evidence for the incremental processing of German quantifier restriction. *Language and Cognition*, *9*, 603–636. <https://doi.org/10.1017/langcog.2016.30>
- 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). <https://doi.org/10.1111/cogs.2019.43.issue-8>
- Augurzky, P., Hohaus, V., & Ulrich, R. (2020). Context and complexity in incremental sentence interpretation: An ERP study on temporal quantification. *Cognitive Science*, *44*(11). <https://doi.org/10.1111/cogs.v44.11>
- Augurzky, P., Schlotterbeck, F., & Ulrich, R. (2020). Most (but not all) quantifiers are interpreted immediately in visual context. *Language, Cognition and Neuroscience*, *35*(9), 1203–1222. <https://doi.org/10.1080/23273798.2020.1722846>
- 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. <https://doi.org/10.1162/jocn.1996.8.5.453>
- Bach, E., Jelinek, E., Kratzer, A., & Partee, B. H. (Eds.). (1995). *Quantification in natural languages*. Kluwer Academic Publishers.
- Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29. <https://doi.org/10.1146/psych.2012.63.issue-1>
- Baggio, G. (2008). Processing temporal constraints: An ERP study. *Language Learning*, *58*(S1), 35–55. <https://doi.org/10.1111/lang.2008.58.issue-s1>
- Baggio, G. (2018). *Meaning in the brain*. 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). Palgrave Macmillan.
- Baggio, G., Choma, T., van Lambalgen, M., & Hagoort, P. (2010). Coercion and compositionality. *Journal of Cognitive Neuroscience*, *22*, 2131–2140. <https://doi.org/10.1162/jocn.2009.21303>
- 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 University Press. <https://doi.org/10.1017/CBO9781139236157>
- Baggio, G., van Lambalgen, M., & Hagoort, P. (2008). Computing and recomputing discourse models: An ERP study. *Journal of Memory and Language*, *59*, 36–53. <https://doi.org/10.1016/j.jml.2008.02.005>
- Baggio, G., van Lambalgen, M., & Hagoort, P. (2015). Logic as mar's computational level: Four case studies. *Topics in Cognitive Science*, *7*(2), 287–298. <https://doi.org/10.1111/tops.12125>
- Bailey, K., Mlynarczyk, G., & West, R. (2016). Slow wave activity related to working memory maintenance in the N-Back task. *Journal of Psychophysiology*, *30*, 141–154. <https://doi.org/10.1027/0269-8803/a000164>
- Barriga-Paulino, C. I., Rodríguez-Martínez, E. I., Rojas-Benjumea, M. Á., & Gómez, C. M. (2014). Slow wave maturation on a visual working memory task. *Brain and Cognition*, *88*, 43–54. <https://doi.org/10.1016/j.bandc.2014.04.003>

- Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4, 159–219. <https://doi.org/10.1007/BF00350139>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bird, A. (2021). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*, 74, 965–993. <https://doi.org/10.1093/bjps/axy051>
- Bremnes, H. S., Szymanik, J., & Baggio, G. (2022). Computational complexity explains neural differences in quantifier verification. *Cognition*, 223, Article 105013. <https://doi.org/10.1016/j.cognition.2022.105013>
- Bremnes, H. S., Szymanik, J., & Baggio, G. (2023). Data for 'The interplay of computational complexity and memory load during quantifier verification'. <https://doi.org/10.18710/LODKEF>.
- Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, 10, 12–21. <https://doi.org/10.1080/17470215808416249>
- Carcassi, F., Steinert-Threlkeld, S., & Szymanik, J. (2021). Monotone quantifiers emerge via iterated learning. *Cognitive Science*, 45. <https://doi.org/10.1111/cogs.v45.8>
- Chang, T. M. (1986). Semantic memory: Facts and models. *Psychological Bulletin*, 99, 199–220. <https://doi.org/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). <https://doi.org/10.1073/pnas.1907023116>
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2, 113–124. <https://doi.org/10.1109/TIT.1956.1056813>
- Clark, H. H. (1976). *Semantics and comprehension*. Mouton.
- Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, 3, 472–517. [https://doi.org/10.1016/0010-0285\(72\)90019-9](https://doi.org/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. <https://doi.org/10.3758/BF03197499>
- Coco, M. I., Araujo, S., & Petersson, K. M. (2017). Disentangling stimulus plausibility and contextual congruency: Electrophysiological evidence for differential cognitive dynamics. *Neuropsychologia*, 96, 150–163. <https://doi.org/10.1016/j.neuropsychologia.2016.12.008>
- Coppock, E. (2019). Quantity superlatives in Germanic, or 'Life on the fault line between adjective and determiner'. *Journal of Germanic Linguistics*, 31, 109–200. <https://doi.org/10.1017/S1470542718000089>
- D'Arcy, R. C. N., Connolly, J. F., & Crocker, S. F. (2000). Latency shifts in the N2b component track phonological deviations in spoken words. *Clinical Neurophysiology*, 111, 40–44. [https://doi.org/10.1016/S1388-2457\(99\)00210-2](https://doi.org/10.1016/S1388-2457(99)00210-2)
- Dehaene, S. (2011). *The number sense: How the mind creates mathematics* (2nd ed.). Oxford University Press.
- 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 Linguistic Society.
- 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. <https://doi.org/10.1080/23273798.2014.980750>
- Eriksen, B. A., & 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. <https://doi.org/10.3758/BF03203267>
- Feigenson, L. (2005). A double-dissociation in infants' representations of object arrays. *Cognition*, 95, B37–B48. <https://doi.org/10.1016/j.cognition.2004.07.006>
- 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. <https://doi.org/10.1023/A:1010447102554>
- Fiebach, C. J., Schlesewsky, M., & 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. [https://doi.org/10.1016/S0749-596X\(02\)00004-9](https://doi.org/10.1016/S0749-596X(02)00004-9)
- Freunberger, D., & Nieuwland, M. S. (2016). Incremental comprehension of spoken quantifier sentences: Evidence from brain potentials. *Brain Research*, 1646, 475–481. <https://doi.org/10.1016/j.brainres.2016.06.035>
- Fukuda, K., Mance, I., & Vogel, E. K. (2015). α power modulation and event-related slow wave provide dissociable correlates of visual working memory. *Journal of Neuroscience*, 35, 14009–14016. <https://doi.org/10.1523/JNEUROSCI.5003-14.2015>
- 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. [https://doi.org/10.1016/S0168-5597\(97\)00085-3](https://doi.org/10.1016/S0168-5597(97)00085-3)
- Hackl, M. (2009). On the grammar and processing of proportional quantifiers: Most versus more than half. *Natural Language Semantics*, 17, 63–98. <https://doi.org/10.1007/s11050-008-9039-x>
- Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience*, 15, 883–899. <https://doi.org/10.1162/089892903322370807>
- Harker, K. T., & Connolly, J. F. (2007). Assessment of visual working memory using event-related potentials. *Clinical Neurophysiology*, 118, 2479–2488. <https://doi.org/10.1016/j.clinph.2007.07.026>
- Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages, and computation*. Addison-Wesley.
- Horn, L. R. (1972). *On the semantic properties of logical operators in english* [PhD thesis]. UCLA.
- 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. <https://doi.org/10.3389/fnhum.2019.00291>
- Hunter, T., & Lidz, J. (2013). Conservativity and learnability of determiners. *Journal of Semantics*, 30, 315–334. <https://doi.org/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. <https://doi.org/10.1007/s11050-016-9130-7>
- Hunt III, L., 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. <https://doi.org/10.1016/j.neulet.2012.11.044>
- 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). Springer.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge University Press.
- Just, M. A. (1974). Comprehending quantified sentences: The relation between sentence-picture and semantic memory verification. *Cognitive Psychology*, 6, 216–236. [https://doi.org/10.1016/0010-0285\(74\)90011-5](https://doi.org/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. [https://doi.org/10.1016/S0022-5371\(71\)80051-8](https://doi.org/10.1016/S0022-5371(71)80051-8)
- 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., & Paperno, D. (2017). Overview. In D. Paperno & E. Keenan (Eds.), *Handbook of quantifiers in natural language: Volume II* (pp. 995–1004). Springer.
- Keenan, E., & Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy*, 9, 253–326. <https://doi.org/10.1007/BF00630273>
- 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. <https://doi.org/10.1162/jocn.1995.7.3.376>
- 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. <https://doi.org/10.1162/jocn.1993.5.2.196>
- 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. <https://doi.org/10.1111/psyp.2011.48.issue-4>
- Knoeferle, P., Urbach, T. P., & 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. <https://doi.org/10.1016/j.actpsy.2014.08.004>
- 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. <https://doi.org/10.1111/nyas.14618>
- 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. <https://doi.org/10.1037/0096-3445.121.4.459>
- 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. <https://doi.org/10.1080/23273798.2015.1130233>
- Kusak, G., Grune, K., Hagedorf, 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. [https://doi.org/10.1016/S0167-8760\(00\)00116-1](https://doi.org/10.1016/S0167-8760(00)00116-1)
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Lefebvre, C., Vachon, F., Grimault, S., Thibault, J., Guimond, S., Peretz, I., R. J. Zatorre, & Jolicœur, P. (2013). Distinct electrophysiological indices of maintenance in auditory and visual short-term memory. *Neuropsychologia*, 51, 2939–2952. <https://doi.org/10.1016/j.neuropsychologia.2013.08.003>
- 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. <https://doi.org/10.1016/j.clinph.2005.03.015>
- Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.
- Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. The MIT Press.
- Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*, 44(1), 27–46. <https://doi.org/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. <https://doi.org/10.1007/s11050-010-9062-6>
- Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). The MIT Press.
- Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience & Biobehavioral Reviews*, 62, 100–108. <https://doi.org/10.1016/j.neubiorev.2016.01.003>
- Marchand, Y., D'Arcy, R. C. N., & Connolly, J. F. (2002). Linking neurophysiological and neuropsychological measures for aphasia assessment. *Clinical Neurophysiology*, 113, 1715–1722. [https://doi.org/10.1016/S1388-2457\(02\)00224-9](https://doi.org/10.1016/S1388-2457(02)00224-9)
- 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. <https://doi.org/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. <https://doi.org/10.1016/j.jneumeth.2007.03.024>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.
- Matthewson, L. (2001). Quantification and the nature of cross-linguistic variation. *Natural Language Semantics*, 9, 145–189. <https://doi.org/10.1023/A:1012492911285>
- 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. <https://doi.org/10.1093/cercor/8.7.563>

- Mostowski, M. (1998). Computational semantics for monadic quantifiers. *Journal of Applied Non-Classical Logics*, 8, 107–121. <https://doi.org/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. [https://doi.org/10.1016/S0926-6410\(96\)00070-5](https://doi.org/10.1016/S0926-6410(96)00070-5)
- Münste, T. F., Schiltz, K., & Kutas, M. (1998). When temporal terms belie conceptual order. *Nature*, 395, 71–73. <https://doi.org/10.1038/25731>
- Neath, I., Saint-Aubin, J., Bireta, T. J., Gabel, A. J., Hudson, C. G., & A. M. Surprenant (2019). Short- and long-term memory tasks predict working memory performance, and vice versa. *Canadian Journal of Experimental Psychology*, 73, 79–93. <https://doi.org/10.1037/cep0000157>
- 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. <https://doi.org/10.1037/xlm0000173>
- Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, 85(2), 203–210. [https://doi.org/10.1016/S0093-934X\(03\)00053-1](https://doi.org/10.1016/S0093-934X(03)00053-1)
- Odic, D., & Starr, A. (2018). An introduction to the approximate number system. *Child Development Perspectives*, 12, 223–229. <https://doi.org/10.1111/cdep.2018.12.issue-4>
- 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. <https://doi.org/10.1155/2011/156869>
- Partee, B. H. (2013). The starring role of quantifiers in the history of formal semantics. In V. Punčochář, & P. Šarný (Eds.), *The logica yearbook 2012* (pp. 113–136). College Publications.
- 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. [https://doi.org/10.1016/0013-4694\(94\)00163-F](https://doi.org/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. [https://doi.org/10.1016/S0926-6410\(97\)00032-3](https://doi.org/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. [https://doi.org/10.1016/0168-5597\(92\)90087-R](https://doi.org/10.1016/0168-5597(92)90087-R)
- Peters, S., & Westerståhl, D. (2006). *Quantifiers in language and logic*. Oxford University Press.
- Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, 58, 193–198. <https://doi.org/10.1037/h0049234>
- Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of 'Most': Semantics, numerosity and psychology. *Mind and Language*, 24, 554–585. <https://doi.org/10.1111/mila.2009.24.issue-5>
- 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). Brill.
- Pijnacker, J., Geurts, B., van Lambalgen, M., Buitelaar, J., & Hagoort, P. (2011). Reasoning with exceptions: An event-related brain potentials study. *Journal of Cognitive Neuroscience*, 23, 471–480. <https://doi.org/10.1162/jocn.2009.21360>
- 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. <https://doi.org/10.1016/j.brainres.2012.10.042>
- Quinlan, J. A., Neath, I., & Surprenant, A. M. (2015). Positional uncertainty in the Brown-Peterson paradigm. *Canadian Journal of Experimental Psychology*, 69, 64–71. <https://doi.org/10.1037/cep0000038>
- Rai, M. K., & 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. <https://doi.org/10.1080/00221325.2011.653839>
- 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. <https://doi.org/10.1016/j.neuropsychologia.2016.09.026>
- Rösler, F., Heil, M., & Röder, B. (1997). Slow negative brain potentials as reflections of specific modular resources of cognition. *Biological Psychology*, 45, 109–141. [https://doi.org/10.1016/S0301-0511\(96\)05225-8](https://doi.org/10.1016/S0301-0511(96)05225-8)
- Ruchkin, D. S., Grafman, J., Cameron, K., & Berndt, R. S. (2003). Working memory retention systems: A state of activated long-term memory. *Behavioral and Brain Sciences*, 26, 709–728. <https://doi.org/10.1017/S0140525X03000165>
- 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. [https://doi.org/10.1016/0013-4694\(90\)90096-3](https://doi.org/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. [https://doi.org/10.1016/0926-6410\(92\)90005-C](https://doi.org/10.1016/0926-6410(92)90005-C)
- Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *TRENDS in Cognitive Sciences*, 11, 251–257. <https://doi.org/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. <https://doi.org/10.1038/33396>
- 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. <https://doi.org/10.1037/xlm0000644>
- 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. <https://doi.org/10.1080/23273798.2016.1161806>
- Steinert-Threlkeld, S., & Szymanik, J. (2019). Learnability and semantic universals. *Semantics and Pragmatics*, 12. <https://doi.org/10.3765/sp.12.4>
- Sternberg, S. (1966). High-Speed scanning in human memory. *Science (New York, N.Y.)*, 153, 652–654. <https://doi.org/10.1126/science.153.3736.652>
- Szymanik, J. (2016). *Quantifiers and cognition: Logical and computational perspectives*. Springer.
- Szymanik, J., & Zajenkowski, M. (2010). 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). Springer Verlag.
- Szymanik, J., & Zajenkowski, M. (2011). Contribution of working memory in parity and proportional judgments. *Belgian Journal of Linguistics*, 25, 176–194. <https://doi.org/10.1075/bjl>
- 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).
- 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).
- 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. <https://doi.org/10.1016/j.jml.2015.03.010>
- Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, 63(2), 158–179. <https://doi.org/10.1016/j.jml.2010.03.008>
- van Benthem, J. (1986). *Essays in logical semantics*. D. Reidel Publishing Company.
- 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. <https://doi.org/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. <https://doi.org/10.1111/psyp.2003.40.issue-2>
- van de Pol, I., Lodder, P., van Maanen, L., Steinert-Threlkeld, S., & Szymanik, J. (2023). Quantifiers satisfying semantic universals have shorter minimal description length. *Cognition*, 232, Article 105150. <https://doi.org/10.1016/j.cognition.2022.105150>
- van Lambalgen, M., & Hamm, F. (2005). *The proper treatment of events*. 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. <https://doi.org/10.1037/0278-7393.25.2.394>
- van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science*, 32, 939–984. <https://doi.org/10.1080/03640210801897856>
- van Rooij, I., & Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychological Inquiry*, 31(4), 321–325. <https://doi.org/10.1080/1047840X.2020.1853477>
- van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, 682–697. <https://doi.org/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 University Press.
- Vissers, C. T. W. M., Kolk, H. K. J., van de Meerendonk, N., & D. J. Chwilla (2008). Monitoring in language perception: Evidence from ERPs in a picture–sentence matching task. *Neuropsychologia*, 46, 967–982. <https://doi.org/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. <https://doi.org/10.1038/nature02447>
- 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. <https://doi.org/10.1111/psyp.2001.38.issue-1>
- Wassenaar, M., & Hagoort, P. (2007). Thematic role assignment in patients with Broca's aphasia: Sentence–picture matching electrified. *Neuropsychologia*, 45, 716–740. <https://doi.org/10.1016/j.neuropsychologia.2006.08.016>
- 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. <https://doi.org/10.1038/s41598-019-45880-y>
- 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. <https://doi.org/10.1016/j.jcomdis.2011.07.005>
- 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. <https://doi.org/10.1016/j.intell.2013.06.020>
- Zajenkowski, M., Szymanik, J., & Garraffa, M. (2014). Working memory mechanism in proportional quantifier verification. *Journal of Psycholinguistic Research*, 43(6), 839–853. <https://doi.org/10.1007/s10936-013-9281-3>
- Zwaan, R. A. (2015). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, 23, 1028–1034. <https://doi.org/10.3758/s13423-015-0864-x>
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123, 162–185. <https://doi.org/10.1037/0033-2909.123.2.162>