# Working Memory Contributions of Generalized Quantifiers in Encoding and Verification: A Pupillometry Study

Aniello De Santo[1] & John E. Drury[2]

[1]*Department of Linguistics, Stony Brook University*
[2]*Department of English Literature and Linguistics,Qatar University*

**Abstract**

In the study of generalized quantifiers, it is essential to have an insightful theory of how their meaning is computed. In particular, a sound computational model of quantifier verification should provide a theoretical framework in which to understand the cognitive requirements that have been reportedly associated to different quantified sentences. A large amount of literature has shown that the type of quantifiers used in a sentence significantly affect the verification procedure employed to arrive at a truth-judgment, with consequences for theories of language, numerical cognition, and memory in general. Interestingly, few studies have explored effects of quantifier type on cognitive load during comprehension, in order to distinguish between quantifier characterization and verification procedures. In this study, we address this distinction by examining the processing of quantified sentences in an auditory/visual verification task. We show quantifier-type influences on working memory usage as measured by variations in pupil size during encoding and verification. Finally, we relate these results to theories of quantifier meaning grounded in the approximate number system, and to previous results on quantifier complexity based on the semantic automata model.

*Keywords:* generalized quantifiers, working memory, semantic automata, pupillometry

## 1. Introduction

Generalized quantifiers have been extensively studied in theoretical and experimental linguistics, since they offer a privileged window into the interface between the specification of the meaning of a sentence, and the cognitive strategies to evaluate its truth value with respect to the outside world (Sanford et al., 1994; Lidz et al., 2011). In particular, the idea that understanding the meaning of a linguistic expression could be linked to an *internalized algorithm* for computing its truth-value in a specific context has been around since Frege (Dummett et al., 1981).

Barwise and Cooper (1981) defined quantifiers as noun phrases that functionally assert some property of a particular set and assign a truth value to it. To assign a truth-value to a sentence like *Every dot is blue* one has to understand the meaning of *every*, and identify the primary property to be related to it (*dots* being *blue*). Psychologists, linguists, and cognitive scientists have proposed a large amount of cognitive models relating quantifiers meaning to numerical processing, visual search, magnitude comparison, as well as to the complexity of the verification strategies (Paterson et al., 2009; Sanford et al., 1996; Pietroski, 2010; Steinert-Threlkeld et al., 2015; Degen and Tanenhaus, 2016).

Most of these ideas are grounded in the Interface Transparency Thesis, according to which the semantic representation of a declarative sentence can be directly mapped into the verification procedure used to determine whether the sentence is true or false (Dummett et al., 1981; Horty, 2007). Thus, since the psychological mechanisms required to transparently implement these semantically equivalent alternative strategies are quite distinct, examining the predictions made for the verification procedures associated to each of these psychological processes can provide decisive evidence for distinguishing them.

In this study, we are particularly interested in understanding the role that the semantic representation of different quantifiers plays in engaging cognitive resources during comprehension and verification.

*Email address:* aniello.desanto@stonybrook.edu (Aniello De Santo[1] & John E. Drury[2])

### 1.1. Number Sense and Verification Strategies

The study of how humans comprehend numerical information (both precise and approximate) has played a crucial role in the investigation of how the meaning of different quantifiers influences the specification of verification strategies. One influential model in numerical cognition — suggested to explain the representation of imprecise cardinalities, and thus the ability to compare quantities without counting — is the Approximate Number System (ANS; Dehaene (1999)). This is supposed to be an evolutionarily cognitive resource that generates representations of numerosity across multiple modalities (e.g. visual objects, auditory beeps, a.o.), and develops in human infants — as well as in many nonverbal animals — without need of explicit training (Feigenson et al., 2004). Several works have explored the idea that quantifier comprehension can be conceptualized with the aid of numerical comparison rooted in the ANS (Dehaene, 1999; Halberda and Feigenson, 2008; Pica et al., 2004), suggesting that children are capable of activating the ANS to comprehend quantifiers from early age, and that they learn how to master the interface between the semantics of quantifiers and more precise quantity representations as they grow older (Sarnecka, 2014).

Building on these assumptions, much work has been done in understanding whether the verification strategies used for quantifier comprehension can be explained in terms of cardinality comparison, with no need for precise counting. For instance, Pietroski et al. (2009) used psychophysical methods to adjudicate between hypotheses about *most* that are considered equivalent by standard semantic tests. Assuming that *most* can either be described in terms of a comparison relation between the cardinalities of two sets (*approximation of cardinalities* strategy), or in terms of a correspondence relation between the individual elements of those sets (*one-to-one with remainder* strategy), this study was interested in understanding which description corresponds to the mental representation of a competent speaker of English. In this experiment, adults evaluated *Most of the dots are yellow* as true or false, during a visual display task which manipulated the ease of using one strategy over the other. According to the authors, their results showed that *most* is understood in terms of cardinality comparison even when counting is impossible, thus suggesting a strong tie between quantifier comprehension and the ANS. In a follow up study, Lidz et al. (2011) examined whether performance changes as a function of increasing the diversity of items in the contrast set. They developed an experiment examining adult judging the sentence *Most of the dots are blue* while being showed displays of colored dots, varying both the number and the color of dots targeted by the sentence. They also varied the ratios of blue to non-blue dots, with half the trials containing more blue dots and half containing more non-blue ones. Again, results indicated use of the ANS in verification, with accuracy unaffected by the number of colors.

Both Pietroski et al. (2009) and Lidz et al. (2011) concluded that semantic judgments are therefore driven by algorithms that transparently compute the relation expressed in the meaning, even when there is a more precise algorithm that is native to the interface system. These results show that the semantic representation of a quantifier (e.g. the canonical specification of *most* in Lidz et al. (2011)) plays a determinative role in identifying the corresponding verification procedure, at least when a transparent strategy is available.

Similar approaches have specifically targeted the contributions of the underlying numerical and linguistic knowledge involved in quantifier comprehension (Heim et al., 2012; Shikhare et al., 2015; Heim et al., 2016). For instance, Shikhare et al. (2015) argue that numerical estimation is crucial in evaluating quantifier sentences under time pressure and propose that numerosity comparison depends on a fixed, externally specified numerical reference for numerical quantifiers (e.g. *at least 3*), while it relies on an internal numerical criterion for proportional quantifiers (e.g. *more than half*).

In sum, there is strong evidence for the fact that aspects of cognition like the ANS enforce constraints on the representational vocabulary of the lexicon itself, particularly when it comes to the implicit representation of generalized quantifiers, and to the evaluation procedures involved in understanding different types of quantifiers. Moreover, these results highlight how there seem to be verification procedures that are more costly (in terms of cognitive resources) than others.

### 1.2. Quantifier Meaning and Computational Complexity

In order to account for the variability among verification procedures associated to different quantifiers, most studies have relied on the predictions made by separating quantifiers in different logic classes.

It is well-known that generalized quantifiers can be defined with different kinds of logics, thus allowing semanticists to classify them based on standard logic hierarchies. As an example, quantifiers like *all, some* are first-order definable, while *most* or *an even number of* are higher-order definable.

However, it doesn't seem that the simple separation between first-order and higher-order logics can give us

the right level of granularity to explain processing differences between different kinds of quantifiers as reported in recent literature (Szymanik and Zajenkowski, 2010). Moreover, the link between logics and verification procedures, and especially demands on cognitive resources, is far from obvious. This calls for a computational theory of quantifiers' complexity with a more transparent mapping to processing and cognitive requirements. Following these ideas, a different line of work on the cognitive systems underlying the comprehension of quantifier meaning has been grounded in the computational model of Van Benthem (1986); which associates quantifiers to computational mechanisms (automata) implementing specific recognition procedures employed for the verification process, via an algorithmic approach based on counting.

The essential intuition behind this model is that the more complex the automaton, the longer the reaction time and working memory involvement of subjects asked to solve the verification task. If we sort quantified expressions in the following groups (Clark and Grossman, 2007):

- Aristotelian: *all, every, some, no, ...*
- Numerical: *at least three, at most four, between eight and ten, ...*
- Parity: *a even number, an odd number*
- Proportional: *most, more than half, ...*

an automata characterization then predicts the following complexity hierarchy: *Aristotelian* $<$ *Parity* $<$ *Numerical* $<$ *Proportional*.

The *semantic automata* framework thus places quantifiers in a hierarchy based on the complexity of the machines required for their verification, making it possible to test the predictions of the model with psycholinguistic and neurolinguistic experiments targeting processing times and memory activation[1] (Szymanik, 2016).

Szymanik and Zajenkowski (2010) used processing experiments based on a verification task over a visual scene to explore the complexity hierarchy as predicted by the automata model. They produced behavioral results matching the model predictions, with sentences containing aristotelian quantifiers like *some* corresponding to the shortest response times, and sentences containing proportional quantifiers such as *more than half* corresponding to the longest reaction times. These results were argued to reflect a higher engagement of working memory when processing proportional quantifiers than aristotelian ones.

Szymanik and Zajenkowski (2010) also explored the role of numbers in the numerical quantifier, showing that these quantifiers are positively correlated with the number they explicitly refer to: the higher the magnitude, the harder is the processing (e.g. *less than eight* vs. *less than four*). Furthermore, when a memory task was used in conjunction with a quantifier verification task, the numerosity in the quantifier expression once again acted as a predictor of the cognitive load for understanding numerical quantifiers (Zajenkowski et al., 2011, 2013). Moreover, distance effects were observed (one vs. three), which were explained in terms of the integration process required for numerical comparison.

In synthesis, there is a significant amount of evidence supporting the semantic automata framework as a good model of specific cognitive aspect of the semantics for generalized quantifiers. The empirical data shows that the computational distinctions made by the hierarchy of automata are in fact reflected in human quantifier processing.

Notably — and in contrast with the assumptions made by ANS-based accounts — the verification algorithms specified by the semantic automata model seem to always rely on precise counting. Interestingly though, results emphasizing the difference between approximate vs. precise judgments were not reported for any comparison across quantifier types in studies probing the cognitive predictions made by semantic automata (Szymanik and Zajenkowski, 2011). Interested in how the verification of generalized quantifiers interacts with (precise vs approximate) number sense, and in how this fits with the automata characterization, Shikhare et al. (2015) explored how adult participants evaluate auditorily presented numerical and proportional quantifier sentences about visually presented numerosities. They showed that participants used numerical estimation and comparison strategies that were biased by the quantifier semantics, and that numerical estimation seems to play an essential role in evaluating quantifier sentences under time pressure.

Therefore, while it appears that the semantic automata model makes the right predictions in terms of

[1] More recently, Graf (2017) has discussed the relation between quantifiers and automata with respect to the computational power of classes of formal languages in the sub-regular hierarchy (Heinz and Idsardi, 2013). This characterization makes finer-grained complexity predictions than the semantic automata model. Moreover, since it directly relies on Van Benthem (1986)'s original definition of quantifier languages and allows us to draw distinctions between quantifiers without reference to a specific recognition mechanism, it might be more appropriate to get insight about complexity distinctions during comprehension.

processing complexity of quantifiers, significant work remains to be done in order to obtain a complete picture of the relationship between truth-conditions, numerical estimation, verification strategies, and memory load.

### *1.3. Current Study*

All the studies mentioned above show differences in the processing of quantifiers influenced by the verification strategies adopted to arrive at a truth-value judgment. In turn, these strategies have been extensively argued to depend on the semantic specification of the quantified expression. Curiously, while the amount of work focusing of differences among quantifiers in terms of verification procedure is overwhelming, almost nothing has been done in the attempt to probe cognitive distinctions during comprehension alone.

In fact, if it is true that latest results present evidence for a link between representations of truth-conditions and verification, it is also evident that studying verification tasks alone can provide only *some* information about comprehension, but not all of it. For instance, in the case of comparative versus superlative quantifiers, it has been observed that people might use similar verification strategies but the process of comprehension might be more complex for superlative quantifiers (Dotlacil et al., 2014). In addition, Szymanik and Zajenkowski (2011) suggest that monotonicity effects go in diverging directions with respect to comprehension and verification, depending on the cognitive task.

In this study then, we are interested in probing the effect of processing the meaning of distinct classes of quantifiers on working memory resources, specifically disentangling comprehension (encoding) from verification.

Notably, previous studies relied on response time (RT) to index the amount of memory resources involved by verification procedures of different complexity (more complex = longer verification time = increased working memory involvement). However, the link between RTs and cognitive load can be inaccurate, especially when a visual task is involved (Attar et al., 2016). Particularly, recent results have cast doubt on the fact that search performance (i.e. RTs) can be used as a good estimator of the amount of working memory engaged in a specific task (Anderson et al., 2013; Emrich et al., 2009; Kane et al., 2006, a.o.) and suggest that RTs collected at the point of response might not correspond to the time at which the meaning of a statement is known to a participant, but might be biased by additional processing due to factors specifically related to the search task (Troiani et al., 2009). Furthermore, relying simply on RTs collected at the end of the verification task makes it impossible to distinguish between processing effects due to the verification procedure from those due simply to the encoding of the quantifier.

Thus, we plan to evaluate the cognitive complexity of different generalized quantifiers by using pupillometry: event-related measures of the variations in subjects' pupil size. Many studies have illustrated a correspondence between pupillary dilation and working memory load (Stanners et al., 1979; Laeng et al., 2012; Nuthmann and Van Der Meer, 2005; Karatekin et al., 2004; Ahern and Beatty, 1979). Variations in pupil size have also been widely used as an estimate of working memory in visual search tasks (Just et al., 2003), and have been shown to be sensitive to local resource demands imposed by sentence comprehension (Engelhardt et al., 2010).

Pupillometry seems then to be a privileged technique to probe working memory demands as associated to the comprehension of quantified expressions. Consistently with the main contrasts explored in previous studies, we selected quantifiers from four different categories (aristotelian, proportional, numerical, cardinal) and exploited pupillometry measures to address two questions:

- are there any effects of quantifier types on working memory specifically during comprehension, before subjects are allowed to engage in verification?

- if effects on memory are found, do they pattern as predicted by computational accounts of quantifier complexity?

The experimental design we used to address these questions is shown in Figure 1. Participants were asked to judge auditory stimulus sentences of the type <*Quantifier*> *of the dots are* <*Color*>, against a visual display showing systematically varied proportions of two sets of colored dots. For numerical quantifiers, the numerical referents were varied systematically in order to probe cardinality effects on pupil size and response time. Crucially, the onset of the visual display was delayed until the onset of the disambiguating predicate, to allow us to measure increases in pupil size relative to each quantifier during *encoding* — prior to any disambiguating or search cue (e.g. the color predicate; the visual scene) — and during *verification*. Proportions of colors in the visual arrays were varied so to avoid fixed counting strategies. Differently from previous studies and to avoid approximation strategies promoted by external time constraints, participants were allowed to provide a response at any time after the presentation of the visual information.

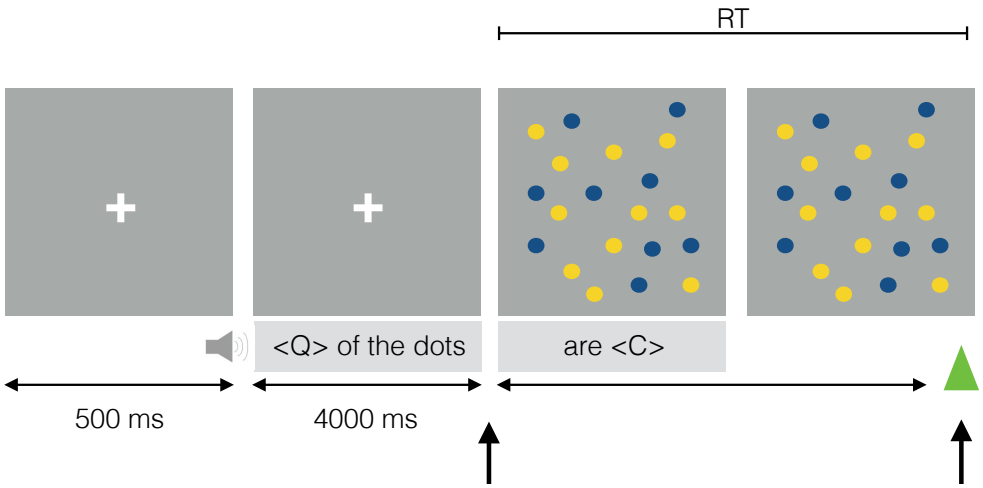



Figure 1: Experimental design.

## 2. Methods

### 2.1. Apparatus

Eye movements and pupil area were recorded using an SR Research EyeLink 1000 desktop system using 35mm lens, at a sampling frequency of 500Hz. After calibration, the average calibration error was 0.5°. Stimuli were presented on an iMac (21.5-inch (diagonal) LED-backlit display with IPS technology; $1920 \times 1080$ resolution; 60Hz refresh rate). All viewers sat at a distance of approximately 90cm from the screen in a room with a dim light setup and used a chin rest to stabilize their head. The camera itself was 60cm away from the eyes, so 30cm forward from the screen. Only the right eye was tracked. The experiment was designed and presented using SR Research Experiment Builder.

### 2.2. Participants

All participants signed consent forms approved by Stony Brook University institutional review board (IRB). A total of 21 healthy adults (age: $20-35$; male:4; female:17) participated in the study in exchange for extra credits. All were right-handed native English speakers with normal or corrected-to-normal vision. Of these participants, 2 were excluded for failure to complete all trials (two blocks out of four), and 2 were excluded for substantial pupil-loss due to blinks (see Sec. 2.5) or inaccurate eyetracking calibration. Accuracy for the whole task was expected to reach a minimum of at least 85%. All participant fulfilled this criterion. Thus, 17 participants (male:3; female:14) were included in the final analyses.

### 2.3. Procedure

As mentioned in Section 1.3, participants were asked to judge the truth-value of auditory stimulus sentences of the type *<Quantifier> of the dots are <Color>*, against a visual display showing systematically varied proportions of two sets of colored dots. The experiment was divided in five parts: a short practice session (4 trials) followed by four experimental blocks. At the beginning of each block, after reading the instructions, a standard 9-point grid calibration and validation of the gaze recording were completed. Since participants were allowed to rest after each block, calibration was repeated after each break, and repeated again at a beginning of a trial in case of noticeable tracking errors. Drift-correct checks were performed before every-trial.

Each trial began with a fixation-cross. After 500ms participants listened to the first auditory phase of an item: *<Quantifier> of the dots*, while the fixation-cross stayed on. In all trials, predicate onset (*are <Color>*) was played exactly 4000ms after quantifier onset. This time window was chosen to allow pupil responses due to the quantifier type to reach their peak (approx. 1200*ms* (Mathôt et al., 2014)) before subjects could engage in verification. The onset of the disambiguating predicate onset was made coincide with the presentation of a visual display with a random distribution of colored circles (*yellow* or *blue*) against a gray background. Subsequently, a blank gray screen was presented for 20ms to allow for blinks and account for screen-refresh time. The same set of auditory stimuli and visual displays was used for all participants in an individually randomized order.

Participants were asked to express their judgment

about the truth-value of the sentence by pressing a key (*f* or *j*, associated with false and true respectively) at any point after the presentation of display. No time constraint was given for the decision phase, but subjects were instructed to react as quickly as possible, and the visual display stayed on until a participant reached a decision. The average length of the whole task was 1 hour.

### 2.4. Materials

We prepared quantified sentences comprising nine quantifiers divided in four main categories: aristotelian (*all, no, some*), proportional (*most, more than half*), numerical (*at least n, at most n*), and parity (*an even number, an odd number*) quantifiers (see Table 1).

Each quantifier was associated to two target colors (*blue*, *yellow*) in two verification conditions (*true*, *false*). Since either of the two colors could be the target color, each quantifier-color combination was presented for 6 trials in *true* condition, and 6 trials in *false* condition. Thus, each quantifier was presented 24 times, for a total of 216 trials.

The visual displays consisted of varying yellow and blue dots, and were drawn using Matlab Psychtoolbox. While the total number of dots in the display was kept constant and equal to 16, proportions of blue and yellow dots were systematically varied based on the truth-conditional properties of the associated quantifier for a total of 12 proportions. Dots were randomly distributed across proportions and matched for size (20 pixels).

Luminance for yellow (RGB: 110) and blue (RGB: 001), as well as the background color (grey: identical among fixation-cross, blank resting screen, and dot arrays), was controlled for all images and set at half of the luminance of white.

The raw material for the auditory stimuli was recorded in a single take using a *Shure SM-54* microphone and a *Zoom H6 digital* recorder. Our speaker was a male native speaker of American English in his mid 20s. The constituents of the stimuli (Quantifier, PP, Color predicate) were cross-spliced using Audacity (R) and Pratt (Boersma and Weenink, 2009), and pasted over a sentence template to make sure that the onset of the color predicate was always at a fixed distance from the onset of the quantifier, and that no differences in pitch and intonation were present across quantifiers (Degen and Tanenhaus, 2016).

### 2.5. Data analysis

SR Research DataViewer was used to output trial reports for three distinct interest periods: baseline (0-500ms), encoding (500-4500ms), and verification (4500ms to key-press). Data points corresponding to blinks were filtered out, together with 10 samples before and after the blink (Mathôt et al., 2018).

Data analysis was subsequently carried on in R. Trials were excluded if more than 10% of data points were missing due to blinks, and a participant was excluded if more than 5% of the trials had been filtered out.

For each interest period and each trial, pupil size values exceeding 2 standard deviation (mean $\pm$ SD) were replaced with the mean pupil size value of the associated condition (Mathôt et al., 2014, 2018; Attar et al., 2016). Moreover, incorrect responses were also excluded from the analysis. Finally, mean and max pupil responses for encoding and verification were computed by subtracting mean pupil baseline at each trial from mean and max. pupil size at each sample, and then averaged across subjects and across trials.

Quantifiers were scored individually and by type. Max and mean pupil response were analyzed separately for each interest period (encoding and verification). Trivially, response times were analyzed only for the verification phase, and computed from the onset of the color predicate to button-press.

For each interest period, we fit linear-mixed models with RT or mean/max pupil response as dependent variables, Quantifier Type (4 levels) and Proportion (14 levels) as fixed effects, and Participant as a random effect.

| Quantifier | Magnitude | Quantifier Category |
|---|---|---|
| All | | Aristotelian |
| No | | |
| Some | | |
| At least *n* | $n = 2,\ldots,7;9\ldots14$ | Numerical |
| At most *n* | $n = 2,\ldots,7;9\ldots14$ | |
| An even number of | | Parity |
| An odd number of | | |
| Most | | Proportional |
| More than half | | |

Table 1: Quantifiers grouped by category

| Quantifier Category | Mean Accuracy (%) | SD (%) |
|---|---|---|
| Aristotelian | 97.1 | 16.66 |
| Parity | 94.5 | 22.68 |
| Numerical | 81.3 | 38.93 |
| Proportional | 98.46 | 12.29 |

Table 2: Accuracy Results

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Quantifier Category | 3 | 224.40 | 74.80 | 662.23 | 0.0000 |
| Proportion | 15 | 19.27 | 1.28 | 11.37 | 0.0000 |
| Residuals | 3189 | 360.21 | 0.11 | | |

Table 3: Output of Anova for RT in verification

## 3. Results

### 3.1. Behavioral Results

As expected, the tasks were quite simple and subjects made overall few mistakes (see Table 2). Accuracy was relatively lower for numerical quantifiers compared to other categories, but no significant statistical effect of quantifier category was found. As expected, the linear mixed effects model revealed significant effects on response times both for quantifier category ($F(3,3189) = 662.23$, $p < 0.001$) and proportion ($F(15,3189) = 11.37$, $p < 0.001$).

Post hoc Tukey comparison of means showed faster response times for Aristotelian < Proportional < Parity/Numerical (see Figure 2), with no significant differences between RTs associated to Parity and Numerical quantifiers ($p < 0.986$) (see Appendix A for details).

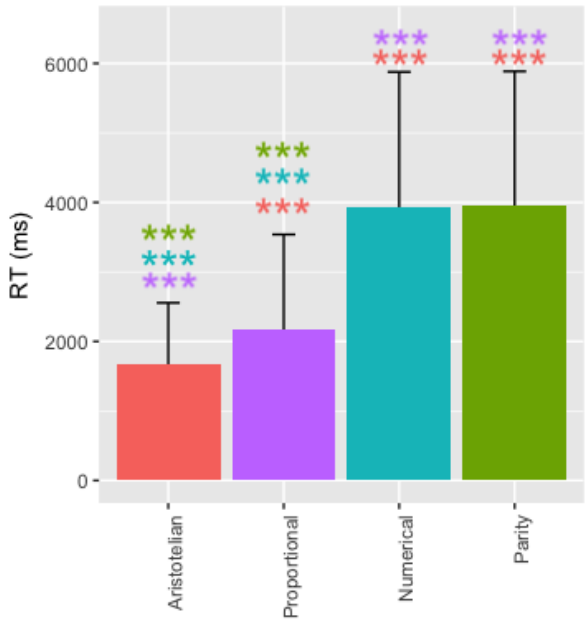



Figure 2: Comparisons of means by quantifier category for RT (in milliseconds) during verification. Signif. codes ($***: 0.001; **: 0.01; *: 0.05$) are color coded based on the quantifier category of reference.

### 3.2. Pupillometry Results

#### 3.2.1. Encoding

The linear mixed effects model and subsequent analysis of variance revealed significant effects of quantifier type on mean ($F(3,3190) = 7.36$, $p < 0.001$) and max ($F(3,3190) = 8.14$, $p < 0.001$) pupil response during the encoding phase, confirming that there were comprehension effects on working memory guided by the semantic content of different quantifiers. As expected, since no visual display was presented in this phase, we found no effects of proportion (mean: $F(15,3190) = 0.86$, $p < 0.611$; max: $F(15,3190) = 0.62$, $p < 0.858$). Post hoc Tukey comparison of means showed that quantifier effects cluster in two main groups, with aristotelian and proportional quantifiers eliciting significantly smaller pupil responses than parity and numerical ones (see Figure 3). No significant differences were found within Aristotelian-Proportional (mean:$p < 0.98$; max:$p < 0.50$) and Parity-Numerical (mean:$p < 0.54$; max: $p < 0.90$) clusters (see Appendix A).

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Quantifier Category | 3 | 0.48 | 0.16 | 7.36 | 0.0001 |
| Proportion | 15 | 0.28 | 0.02 | 0.86 | 0.6114 |
| Residuals | 3190 | 69.78 | 0.02 | | |

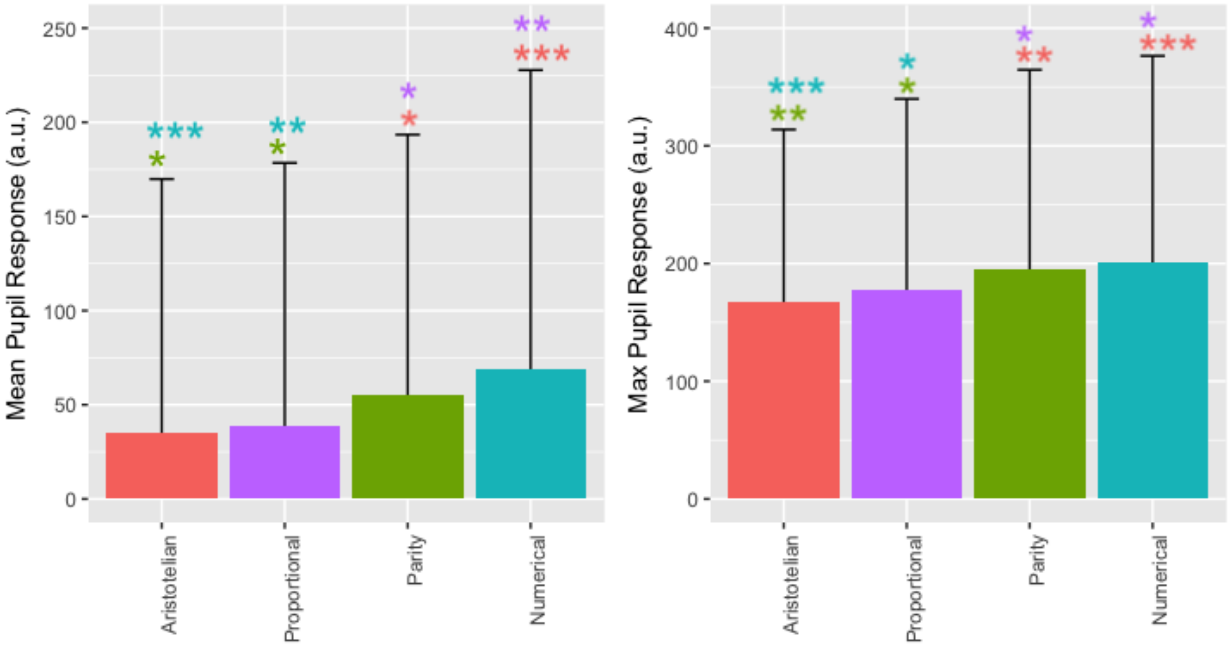
Table 4: Output of Anova for mean pupil response in encoding

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Quantifier Category | 3 | 637454.43 | 212484.81 | 8.14 | 0.0000 |
| Proportion | 15 | 243735.71 | 16249.05 | 0.62 | 0.8587 |
| Residuals | 3190 | 83226450.12 | 26089.80 | | |

Table 5: Output of Anova for max pupil response in encoding

#### 3.2.2. Verification

Significant effects were found of quantifier type on mean ($F(3,3189) = 5.117$, $p < 0.01$) and max ($F(3,3190) = 31.740$, $p < 0.001$) pupil response during verification. Maybe surprisingly, we also found no effects of proportion (mean: $F(15,3190) = 0.218$, $p < 0.611$; max: $F(15,3190) = 1.091$, $p < 0.358$) on either mean nor max pupillary response. Post Tukey comparison of means again showed significantly smaller pupil responses for aristotelian-proportional quantifiers than for parity and numerical quantifiers (see Figure 4), with no significant differences within Aristotelian-Proportional (mean:$p < 0.16$; max: $p < 0.94$) and Parity-Numerical (mean:$p < 0.63$; max: $p < 0.55$) clusters, respectively. Notably, the difference between aristotelian and parity mean pupil response was also not significant ($p < 0.82$) (see Appendix A).




(a) (b)

Figure 3: Comparisons of means by quantifier category for (a) mean and b) max pupil response (in arbitrary units) during encoding. Signif. codes ($***: 0.001; **: 0.01; *: 0.05$) are color coded based on the quantifier category of reference.

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Quantifier Category | 3 | 0.64 | 0.21 | 5.12 | 0.0016 |
| Proportion | 15 | 0.78 | 0.05 | 1.26 | 0.2183 |
| Residuals | 3189 | 132.04 | 0.04 | | |

Table 6: Output of Anova for mean pupil response in verification

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| quant_cat | 3 | 5259016.31 | 1753005.44 | 31.74 | 0.0000 |
| prop | 15 | 904239.19 | 60282.61 | 1.09 | 0.3583 |
| Residuals | 3189 | 176129452.17 | 55230.31 | | |

Table 7: Output of Anova for max pupil response in verification

## 4. Discussion

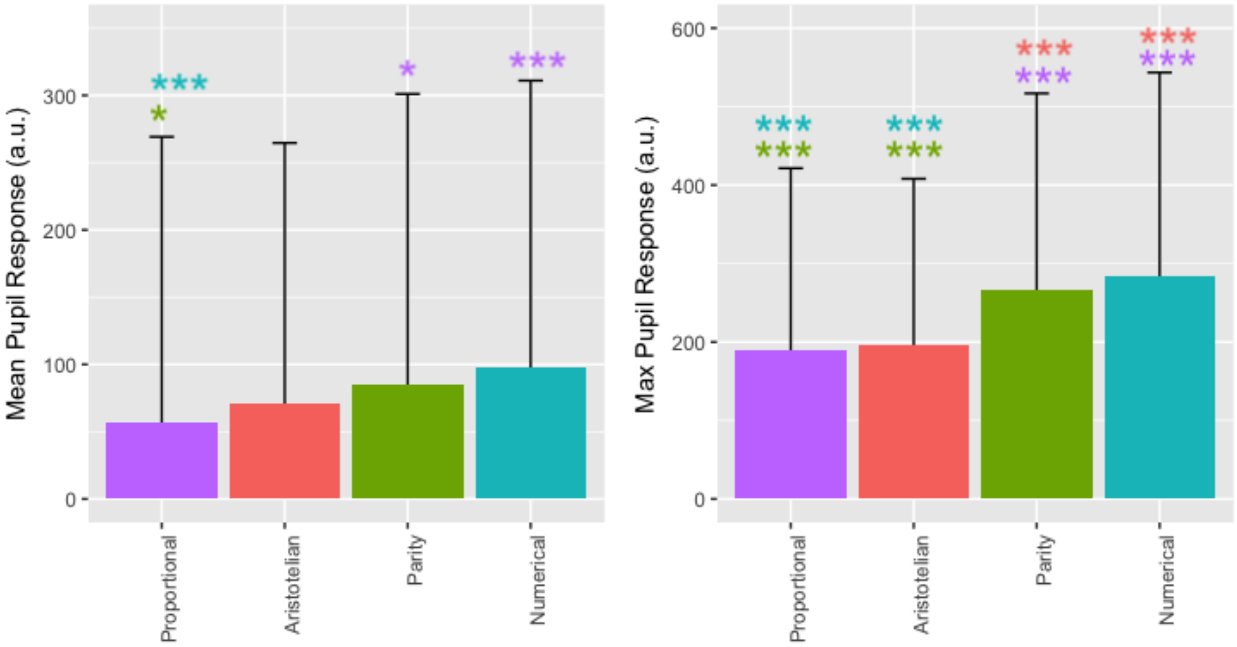
This study employed recordings of pupil size variation during a truth-value judgment task to better understand cognitive resources underlying the processing of quantified sentences. In particular, we were interested in exploring whether effects of different kind of quantifiers (namely, aristotelian, proportional, numerical, and parity) could be found during *encoding*: a phase in which subjects had heard a quantified expression, but had not yet been given access to a disambiguating predicate or a visual scene to contrast the quantifier with. Moreover, we wondered whether quantifier type effects on pupil response both during encoding and verification could be explained by a model of quantifier complexity based on the semantic automata framework.

With respect to the first question, our results show significant effects of quantifier type during the encoding period, indicating that working memory is in fact being modulated by quantifier type even before the subject could engage in any type of verification strategy.

Interestingly, these effects do not seem to mirror the complexity hierarchy proposed by the semantic automata model, with proportional quantifiers patterning together with aristotelian quantifiers in recruiting significantly less resources than numerical and parity (Szymanik, 2016).

These results are in line with the idea that variations in working memory load during this early phase are associated to the encoding of precise numerical concepts, and to initial allocation of cognitive resources that are going to be needed by the different verification procedure associated with different quantifiers' types.

It has been observed that aristotelian quantifiers do not require precise estimations of the cardinalities of the target sets to arrive at a truth-judgment. Thus, they ini-




(a) (b)

Figure 4: Comparisons of means by quantifier category for (a) mean and b) max pupil response (in arbitrary units) during verification. Signif. codes ($***: 0.001; **: 0.01; *: 0.05$) are color coded based on the quantifier category of reference.

tially require relatively small cognitive resources, possibly associated to the need for approximate comparisons.

On the contrary, parity and numerical quantifiers have consistently shown automatic access to specific numerical magnitudes (Troiani et al., 2009; Dehaene and Cohen, 1997). Since these quantifiers always presuppose precise numerical comparisons, it is probable that the increase in pupil responses is indexing the initial recruitment of additional resources needed to retrieve the target numerical representation and actively maintaining it in memory (Dehaene et al., 2003; Heim et al., 2012, 2016). In this perspective, the fact that no differences were found between parity and numerical quantifiers across interest periods should also not be surprising (Troiani et al., 2009).

Finally, if the initial specification of proportional quantifiers relies on approximate comparisons between sets instead of precise one-to-one counting (Dehaene and Cohen, 1997; Pietroski et al., 2009), we would expect the recruitment of resources associated to computing vague numerical concepts with no need for precise magnitude maintenance. It is then expected that the corresponding increase in working memory load as indexed by pupil response would pattern similarly to aristotelian quantifiers, and be smaller than the one associated to numerical/parity quantifiers.

Similar response patterns both for mean pupil response and for max response peak are found during the verification phase. This result, together with the fact that pupil variation was still not significantly affected by the proportions of target colors in the visual scene, suggests that how the verification procedure is carried on for distinct quantifiers plays a less crucial role in modulating cognitive load that what was previously reported.

Again in contrast with what reported by experiments on the semantic automata model, response times also showed a similar pattern: with aristotelian quantifiers being associated to the shortest RTs, and numerical/parity quantifiers to the longest onest.

Interestingly, while RTs for proportional quantifiers overall pattern similarly to pupil responses, they also show significant differences with aristotelian quantifiers. We interpret this apparent mismatch between pupil response and RT as evidence of the fact that the amount of working memory recruited for verification is mostly modulated by quantifier encoding in the initial stages of comprehension, while response times are instead affected by the length of the verification proce-

dures. To verify the meaning of an expression containing an aristotelian quantifier it suffices to identify a single target element; proportional quantifiers are instead going to require approximate cardinalities of large sets, thus leading to longer search over the visual scene.

If this is true, then we also predict that RTs for proportional quantifiers should be longer, the closer the proportions of the target sets are to requiring precise numerical comparisons. Although our design was not meant to conduct proportion-by-proportion comparisons across quantifiers, we can see an effect consistent with this prediction in Figure 5. Here, the RTs associated to proportional quantifiers stick close together with those for aristotelian while the proportions of the target sets are far from each other, but visibly increase towards numerical and parity quantifiers when the proportions of the sets are close to each other.

Overall, we take the fact that response times are sensitive to the complexity of the visual scene — pupil response, neither during encoding or verification, was not — as confirming that response times should not be used as a good indicator of working memory load in tasks that involve language comprehension simultaneously to picture verification. In these contexts, response times are possibly just measuring the length taken by the verification procedure. This leads to an important additional distinction, often overlooked in studies of language complexity: while it is true that longer tasks require longer maintenance of information in memory, this should not be taken to be equivalent to an absolute increase in memory burden (in other words, holding something memory for longer time is not equivalent to recruiting more memory resources at a specific time).

Finally, we want to return to a comparison with the semantic automata model. At a first pass, our results seem to be in direct contrast with the predictions made by this model, and with previous experimental results showing proportional quantifiers to be more complex than arisotelian, numerical, and parity.

However, we believe that these apparently conflicting results could be explained by a few differences between the design used in those experiments and our own. Particularly, previous results on proportional quantifiers in the semantic automata literature had used precise borderline proportions in the visual presentations (e.g. seven targets in a scene with fifteen objects), thus never allowing for approximate comparisons in the verification of proportional meanings. In contrast, we allowed proportional quantifiers to be compared over a range of varying proportions. However, since we were interested in varying the target magnitude while keeping the number of trials to a manageable amount, numerical and parity quantifiers were accompanied by scenes close to the target magnitude (e.g. *at least three* and a scene of four blue dots and twelve yellow dots).

Therefore, while we almost always allowed for approximate comparisons in the verification of proportional quantifiers, numerical and parity quantifiers were always presented in a context that forced for precise counting. As already discussed with respect to Figure 5, it is probable that in a set-up where the verification strategies for proportional, parity, and numerical quantifiers are fixed, and approximation strategies are overall disallowed, response times would again pattern as predicted by the semantic automata model and shown by (Szymanik and Zajenkowski, 2011, 2010; Zajenkowski and Szymanik, 2013, a.o.). On the other hand, when precise counting is discouraged across quantifiers, we predict a replication of this paper's results. This is clearly a direction that needs to be explored in future research.

Crucially though, our results already add to the understanding of the predictions made by the semantic automata model. As we discussed above, the mismatch between pupil response and reaction time suggests that previously reported results employing RTs were not measuring exact working memory demands associated to distinct automata, but instead the time taken by each automata to output a truth-judgment.

This might seem to contradict more recent studies providing additional evidence for the involvement of working memory resources — like storage and maintenance of information — during verification of proportional quantifiers (Zajenkowski et al., 2011, 2013). These results combined the usual sentence verification task with traditional measures of memory load. For example, they reduced the availability of memory resources during the verification task by presenting a numerical sequence to the participants before the visual task, and then asking them to retrieve part of the sequence post task. Crucially, they showed that, while the accuracy of sentences involving aristotelian and numerical quantifiers was unaffected by the memory task, the accuracy of proportional quantifier decreased consistently, thus supporting the idea that the latter are more demanding in terms of memory resources.

As already mentioned though, it has been observed how the direct assessment of working memory load through memory retrieval tasks interferes with the search task, making the results of such experiments difficult to interpret (Attar et al., 2016, a.o.). Instead, in light of our own results, it seems reasonable to suggest that the variation in response times reported by (Zajenkowski et al., 2013, a.o.) reflects the need for

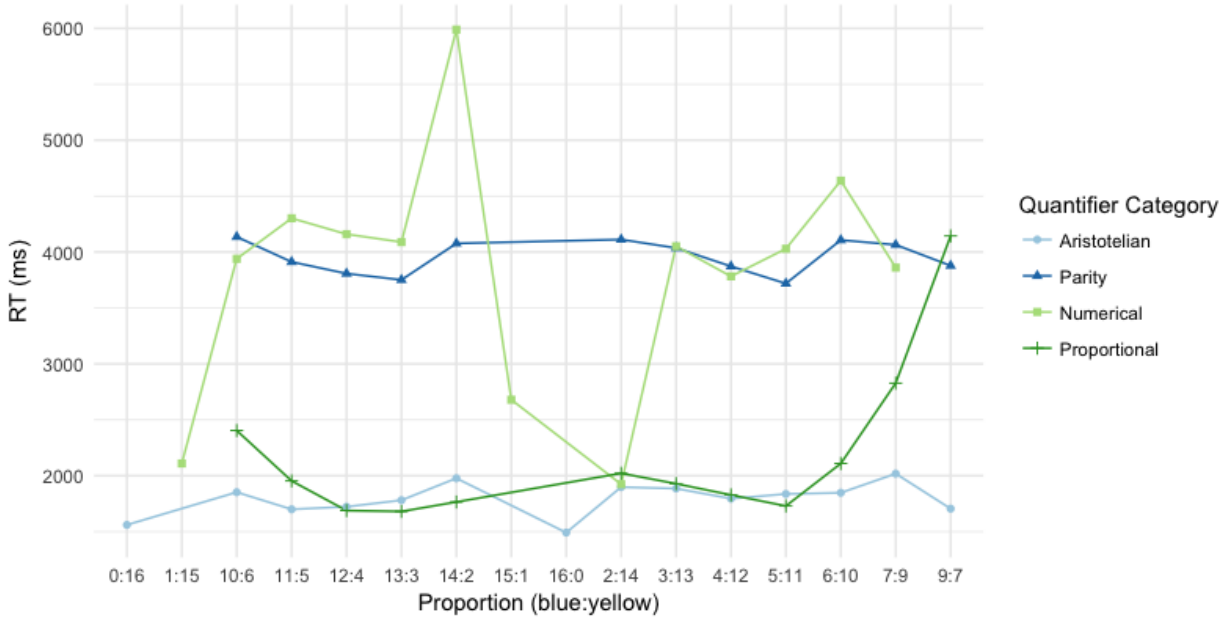



Figure 5: Comparisons of RT by quantifier category (in milliseconds) and proportions of colored dots (*blue:yellow*) during verification.

more complex search strategies for proportional quantifiers that might be affected by a concurrent memory task, but do not directly index additional requirements on working memory.

In sum, the patterns of pupil response and response times emerging in our study seem to be consistent both with the ANS (pupil variation indexing the amount of resources recruited based on the quantifier meaning) and with the semantic automata model (response times indexing the length of different verification procedure).

Finally, one last consideration is worth making about the semantic automata claims with respect to memory and proportional quantifiers. The original characterization of quantifiers in terms of automata shows that aristotelian and numerical quantifiers can be specified by machines called finite-state automata, requiring very simple short term memory mechanisms. Others quantifiers — notably proportional quantifiers like *most* and *more than half* — require for their verification machines called pushdown automata, which essentially augments finite-state automata with a form of memory called *stack* (Van Benthem, 1986; Mostowski, 1957; Hopcroft et al., 2001). Therefore, the intuition exploited in previous experimental studies is that proportional quantifiers have more significant working memory requirements than aristotelian quantifiers, due to the differences in memory mechanisms of the corresponding automata.

However, the automata model does not in fact directly predict that proportional quantifiers should require increased working memory resources. A more direct prediction would be not in the *amount* of memory, but in the *type* of memory recruited by quantifiers associated to distinct automata. This is consistent with neuroimaging work showing that quantified sentences' verification recruits the right inferior parietal cortex — associated with numerosity — independently of the quantifier involved, but that only proportional quantifiers recruit the prefrontal cortex, which is generally associated with executive resources such as working memory (McMillan et al., 2005; Troiani et al., 2009; Heim et al., 2016). Moreover, it aligns with ideas about finite-state and push-down memory structures proposed independently in the literature (Gallistel and Gelman, 2000; Fitch, 2014), and with recent electrophysiological results on quantified sentences' comprehension (De Santo et al., to appear).

To the best of our knowledge, no neuroimaging study contrasted activations during encoding and verification[2]. In future it would be interesting to employ a design similar to ours and see if neuroimaging results produce results comparable to the patterns of pupil response reported in the present study.

[2]Interestingly, the design in Heim et al. (2016) would have allowed for this contrast, but no comparisons between quantifiers during encoding were reported (see aslo (Hackl, 2009)).

## 5. Conclusion

In this study, we used variations in pupil diameter to probe quantifier type influences on workin memory load during encoding and verification of quantified sentences. We showed that different types of quantified expressions modulate different pupillary responses already during comprehension, before any cue to verification has been given. We tied these early responses to the recruitment of cognitive resources used to decide the truth-valued of the quantified expression, due to different verification procedures associated to the type of quantifier. Finally, we compared our results with predictions about quantifier complexity made by the semantic automata model, and proposed a way to reconcile the precise counting strategy this model relies on with ideas about approximate comparisons based on the approximate number system. Building on these results, we argue against using response times as indexes of cognitive load, and propose pupillometry as a more effective way to investigate working memory effects in studies probing natural language meaning thorough visual search tasks. We believe that these results, possibly extended by neuroimaging techniques, will contribute not just to theories of generalized quantifiers, but also to our understanding of memory organization and its interaction with language and numerical cognition.

## References


Ahern, S., Beatty, J., 1979. Pupillary responses during information processing vary with scholastic aptitude test scores. Science 205, 1289–1292.

Anderson, D.E., Vogel, E.K., Awh, E., 2013. Retracted: A common discrete resource for visual working memory and visual search. Psychological Science 24, 929–938. doi:10.1177/0956797612464380. pMID: 23572280.

Attar, N., Schneps, M.H., Pomplun, M., 2016. Working memory load predicts visual search efficiency: Evidence from a novel pupillary response paradigm. Memory & Cognition .

Barwise, J., Cooper, R., 1981. Generalized quantifiers and natural language. Linguistics and Philosophy 4, 159–219.

Boersma, P., Weenink, D., 2009. Praat: doing phonetics by computer (version 5.1.13). URL: http://www.praat.org.

Clark, R., Grossman, M., 2007. Number sense and quantifier interpretation. Topoi 26, 51–62.

De Santo, A., Rawski, J., Yazdani, A., Drury, J.E., to appear. Quantified sentences as a window into prediction and priming: An erp study., in: Proceedings of the 54rd Meeting of the Chicago Linguistic Society (CLS54).

Degen, J., Tanenhaus, M.K., 2016. Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. Cognitive Science 40, 172–201. URL: http://dx.doi.org/10.1111/cogs.12227, doi:10.1111/cogs.12227.

Dehaene, S., 1999. The number sense: How the mind creates mathematics. OUP USA.

Dehaene, S., Cohen, L., 1997. Cerebral pathways for calculation: Double dissociation between rote verbal and quantitative knowledge of arithmetic. Cortex 33, 219–250.

Dehaene, S., Piazza, M., Pinel, P., Cohen, L., 2003. Three parietal circuits for number processing. Cognitive neuropsychology 20, 487–506.

Dotlacil, J., Szymanik, J., Zajenkowski, M., 2014. Probabilistic semantic automata in the verification of quantified statements, in: Bello, P., Guarini, M., McShane, M., Scassellati, B. (Eds.), Proceedings of the 36th Annual Meeting of the Cognitive Science Society, CogSci 2014, Quebec City, Canada, July 23-26, 2014, cognitivesciencesociety.org. URL: https://mindmodeling.org/cogsci2014/papers/512/.

Dummett, M.A., et al., 1981. Frege: Philosophy of language. volume 2. Cambridge Univ Press.

Emrich, S.M., Al-Aidroos, N., Pratt, J., Ferber, S., 2009. Visual search elicits the electrophysiological marker of visual working memory. PLOS ONE 4, 1–9. URL: https://doi.org/10.1371/journal.pone.0008042, doi:10.1371/journal.pone.0008042.

Engelhardt, P.E., Ferreira, F., Patsenko, E.G., 2010. Pupillometry reveals processing load during spoken language comprehension. The Quarterly Journal of Experimental Psychology 63, 639–645.

Feigenson, L., Dehaene, S., Spelke, E., 2004. Core systems of number. Trends in cognitive sciences 8, 307–314.

Fitch, W.T., 2014. Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition. Physics of Life Reviews 11, 329 – 364. URL: http://www.sciencedirect.com/science/article/pii/S157106451400058X, doi:http://dx.doi.org/10.1016/j.plrev.2014.04.005.

Gallistel, C.R., Gelman, R., 2000. Non-verbal numerical cognition: From reals to integers. Trends in cognitive sciences 4, 59–65.

Graf, T., 2017. Subregular morpho-semantics: The expressive limits of monomorphemic quantifiers. Invited talk, December 15, Rutgers University, New Brunswick, NJ.

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

Halberda, J., Feigenson, L., 2008. Developmental change in the acuity of the" number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. Developmental psychology 44, 1457.

Heim, S., Amunts, K., Drai, D., Eickhoff, S., Hautvast, S., Grodzinsky, Y., 2012. The language number interface in the brain: A complex parametric study of quantifiers and quantities. Frontiers in Evolutionary Neuroscience 4, 4. URL: http://journal.frontiersin.org/article/10.3389/fnevo.2012.00004, doi:10.3389/fnevo.2012.00004.

Heim, S., McMillan, C.T., Clark, R., Baehr, L., Ternes, K., Olm, C., Min, N.E., Grossman, M., 2016. How the brain learns how few are "many": An fmri study of the flexibility of quantifier semantics. NeuroImage 125, 45–52. URL: http://dx.doi.org/10.1016/j.neuroimage.2015.10.035, doi:10.1016/j.neuroimage.2015.10.035.

Heinz, J., Idsardi, W., 2013. What complexity differences reveal about domains in language. Topics in Cognitive Science 5, 111–131.

Hopcroft, J.E., Motwani, R., Ullman, J.D., 2001. Introduction to automata theory, languages, and computation. Acm Sigact News 32, 60–65.

Horty, J., 2007. Frege on definitions: A case study of semantic content. Oxford University Press.

Just, M.A., Carpenter, P.A., Miyake, A., 2003. Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related po-

tential studies of brain work, in: Theoretical Issues in Ergonomics Science, pp. 56–88.

Kane, M.J., Poole, B.J., Tuholski, S.W., Engle, R.W., 2006. Working memory capacity and the top-down control of visual search: Exploring the boundaries of" executive attention". Journal of Experimental Psychology: Learning, Memory, and Cognition 32, 749.

Karatekin, C., Couperus, J.W., Marcus, D.J., 2004. Attention allocation in the dual-task paradigm as measured through behavioral and psychophysiological responses. Psychophysiology 41, 175–185.

Laeng, B., Sirois, S., Gredebäck, G., 2012. Pupillometry: a window to the preconscious? Perspectives on psychological science 7, 18–27.

Lidz, J., Pietroski, P., Halberda, J., Hunter, T., 2011. Interface transparency and the psychosemantics of most. Natural Language Semantics 19, 227–256. URL: http://dx.doi.org/10.1007/s11050-010-9062-6, doi:10.1007/s11050-010-9062-6.

Mathôt, S., Dalmaijer, E., Grainger, J., Van der Stigchel, S., 2014. The pupillary light response reflects exogenous attention and inhibition of return. Journal of Vision 14, 7–7.

Mathôt, S., Fabius, J., Van Heusden, E., Van der Stigchel, S., 2018. Safe and sensible preprocessing and baseline correction of pupil-size data. Behavior research methods , 1–13.

McMillan, C.T., Clark, R., Moore, P., Devita, C., Grossman, M., 2005. Neural basis for generalized quantifier comprehension. Neuropsychologia 43, 1729–1737.

Mostowski, A., 1957. On a generalization of quantifiers .

Nuthmann, A., Van Der Meer, E., 2005. Time's arrow and pupillary response. Psychophysiology 42, 306–317.

Paterson, K.B., Filik, R., Moxey, L.M., 2009. Quantifiers and discourse processing. Language and Linguistics Compass 3, 1390–1402.

Pica, P., Lemer, C., Izard, V., Dehaene, S., 2004. Exact and approximate arithmetic in an amazonian indigene group. Science 306, 499–503.

Pietroski, P., Lidz, J., Hunter, T., Halberda, J., 2009. The meaning of ?most?: Semantics, numerosity and psychology. Mind & Language 24, 554–585.

Pietroski, P.M., 2010. Concepts, meanings and truth: First nature, second nature and hard work. Mind & Language 25, 247–278.

Sanford, A.J., Moxey, L.M., Paterson, K., 1994. Psychological studies of quantifiers. Journal of Semantics 11, 153–170.

Sanford, A.J., Moxey, L.M., Paterson, K.B., 1996. Attentional focusing with quantifiers in production and comprehension. Memory & Cognition 24, 144–155.

Sarnecka, B.W., 2014. On the relation between grammatical number and cardinal numbers in development. Frontiers in Psychology 5. URL: http://dx.doi.org/10.3389/fpsyg.2014.01132, doi:10.3389/fpsyg.2014.01132.

Shikhare, S., Heim, S., Klein, E., Huber, S., Willmes, K., 2015. Processing of numerical and proportional quantifiers. Cognitive Science 39, 1504–1536. URL: http://dx.doi.org/10.1111/cogs.12219, doi:10.1111/cogs.12219.

Stanners, R.F., Coulter, M., Sweet, A.W., Murphy, P., 1979. The pupillary response as an indicator of arousal and cognition. Motivation and Emotion 3, 319–340.

Steinert-Threlkeld, S., Munneke, G.J., Szymanik, J., 2015. Alternative representations in formal semantics: A case study of quantifiers, in: T. Brochhagen, F. Roelofsen, N.T. (Ed.), Proceedings of the 20th Amsterdam Colloquium, pp. 368–377.

Szymanik, J., 2016. Computing simple quantifiers, in: Quantifiers and Cognition: Logical and Computational Perspectives. Springer International Publishing, pp. 41–49.

Szymanik, J., Zajenkowski, M., 2010. Comprehension of simple quantifiers: Empirical evaluation of a computational model. Cognitive Science 34, 521–532.

Szymanik, J., Zajenkowski, M., 2011. Contribution of working memory in parity and proportional judgments. Belgian Journal of Linguistics 25, 176–194.

Troiani, V., Peelle, J.E., Clark, R., Grossman, M., 2009. Is it logical to count on quantifiers? dissociable neural networks underlying numerical and logical quantifiers. Neuropsychologia 47, 104–111.

Van Benthem, J., 1986. Quantifiers. Springer Netherlands, Dordrecht. pp. 25–54. URL: http://dx.doi.org/10.1007/978-94-009-4540-1_2, doi:10.1007/978-94-009-4540-1_2.

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.

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, 456–466.

Zajenkowski, M., Szymanik, J., Garraffa, M., 2013. Working memory mechanism in proportional quantifier verification. Journal of Psycholinguistic Research .

## Appendix A

| | diff | lwr | upr | p adj |
|---|---|---|---|---|
| Parity-Aristotelian | 0.02 | 0.00 | 0.04 | 0.03 |
| Numerical-Aristotelian | 0.03 | 0.01 | 0.05 | 0.00 |
| Proportional-Aristotelian | 0.00 | -0.02 | 0.02 | 0.98 |
| Numerical-Parity | 0.01 | -0.01 | 0.03 | 0.54 |
| Proportional-Parity | -0.02 | -0.04 | 0.00 | 0.03 |
| Proportional-Numerical | -0.03 | -0.05 | -0.01 | 0.00 |

Table 8: Tukey multiple comparisons of means for mean pupil response in encoding

| | diff | lwr | upr | p adj |
|---|---|---|---|---|
| Parity-Aristotelian | 28.53 | 8.69 | 48.38 | 0.00 |
| Numerical-Aristotelian | 34.74 | 13.94 | 55.55 | 0.00 |
| Proportional-Aristotelian | 10.63 | -8.94 | 30.21 | 0.50 |
| Numerical-Parity | 6.21 | -16.55 | 28.97 | 0.90 |
| Proportional-Parity | -17.90 | -39.53 | 3.73 | 0.04 |
| Proportional-Numerical | -24.11 | -46.63 | -1.59 | 0.03 |

Table 9: Tukey multiple comparisons of means for max pupil response in encoding

| | diff | lwr | upr | p adj |
|---|---|---|---|---|
| Parity-Aristotelian | 0.01 | -0.02 | 0.03 | 0.82 |
| Numerical-Aristotelian | 0.02 | -0.00 | 0.05 | 0.14 |
| Proportional-Aristotelian | -0.02 | -0.04 | 0.00 | 0.16 |
| Numerical-Parity | 0.01 | -0.02 | 0.04 | 0.63 |
| Proportional-Parity | -0.03 | -0.06 | -0.00 | 0.04 |
| Proportional-Numerical | -0.04 | -0.07 | -0.01 | 0.00 |

Table 10: Tukey multiple comparisons of means for mean pupil response in verification

| | diff | lwr | upr | p adj |
|---|---|---|---|---|
| Parity-Aristotelian | 70.82 | 41.94 | 99.70 | 0.00 |
| Numerical-Aristotelian | 87.75 | 57.47 | 118.03 | 0.00 |
| Proportional-Aristotelian | -6.36 | -34.85 | 22.12 | 0.94 |
| Numerical-Parity | 16.93 | -16.18 | 50.04 | 0.55 |
| Proportional-Parity | -77.18 | -108.66 | -45.71 | 0.00 |
| Proportional-Numerical | -94.12 | -126.88 | -61.35 | 0.00 |

Table 11: Tukey multiple comparisons of means for max pupil response in verification

| | diff | lwr | upr | p adj |
|---|---|---|---|---|
| Parity-Aristotelian | 0.59 | 0.54 | 0.63 | 0.00 |
| Numerical-Aristotelian | 0.58 | 0.54 | 0.62 | 0.00 |
| Proportional-Aristotelian | 0.14 | 0.10 | 0.18 | 0.00 |
| Numerical-Parity | -0.01 | -0.05 | 0.04 | 0.99 |
| Proportional-Parity | -0.44 | -0.49 | -0.40 | 0.00 |
| Proportional-Numerical | -0.44 | -0.48 | -0.39 | 0.00 |

Table 12: Tukey multiple comparisons of means for RT in verification