Encoding and Verification Effects of Generalized Quantifiers on Working Memory

Aniello De Santo and John E. Drury
Stony Brook University and Jiangsu Normal University

ABSTRACT

A large amount of literature has shown that the type of quantifier used in a sentence significantly affects the verification procedure and the cognitive resources employed to arrive at a truth-judgment. Interestingly, few studies have explored effects of quantifier type on cognitive load during comprehension alone, 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, and we relate these results to theories of quantifier meaning grounded in the approximate number system, and to previous results on quantifier complexity based on precise counting strategies.

1 Introduction

Barwise & Cooper (1981) define generalized quantifiers as noun phrases that functionally assert some property of a particular set and assign a truth value to it. For instance, 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*).

Thus, in building a cognitive theory of quantifiers' interpretation, it is essential to have an insightful theory of how their meaning is computed. Several studies support the idea that the semantic representation of a quantifier (e.g., its canonical specification, cf. Lidz *et al.* (2011)) plays a determinative role in identifying the corresponding verification procedure — at least when a transparent strategy is available. In this perspective, it has been argued that the relation between the truth-conditional properties of generalized quantifiers and the verification strategies said quantifiers are associated with could be better understood by establishing cross-disciplinary links to logic, numerical processing, visual search, and magnitude comparison (Paterson *et al.* 2009; Degen & Tanenhaus 2016; Pietroski 2010; Steinert-Threlkeld *et al.* 2015).

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.

To appear in the Proceedings of CLS 55.

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). 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 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 & Feigenson 2008), 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.

Building on these assumptions, much work has been done to understand whether the verification strategies used for quantifier comprehension can be explained in terms of cardinality comparison, with no need for precise counting. This line of investigation has provided 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 complexity of their evaluation procedures (Pietroski *et al.* 2009; Lidz *et al.* 2011; Heim *et al.* 2012; Heim *et al.* 2016; Shikhare *et al.* 2015). 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, past studies have separated quantifiers in different logic classes. However, the link between logic classes, verification procedures, and demands on cognitive resources is far from obvious. This calls for a computational theory of quantifiers' complexity with a transparent mapping to processing and cognitive requirements. Following these ideas, the *semantic automata* model associates quantifiers to computational mechanisms (automata) implementing specific recognition procedures employed for the verification process, via an algorithmic approach based on counting (Van Benthem 1986).

The essential intuition behind this model is that the more complex the automaton, the longer the reaction time and working memory involvement will be, for subjects asked to solve the verification task. In this sense, quantifiers are ordered based on the complexity of the machines required for their verification. If we sort quantified expressions in the following groups (Clark & 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*. Recent behavioral experiments have reported evidence in support of such a hierarchy (Szymanik & Zajenkowski 2010; Zajenkowski *et al.* 2011; Zajenkowski *et al.* 2013; Steinert-Threlkeld *et al.* 2015). Notably, and in contrast with the assumptions made by ANS-based accounts, the verification algorithms specified by this model always rely on precise counting.

Interested in how the verification of generalized quantifiers interacts with (precise vs approximate) number sense, Shikhare *et al.* (2015) showed that adults use numerical estimation and comparison strategies 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 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.

Curiously, while the amount of work focusing on how differences among quantifiers affect verification procedures is extensive, few studies have probed cognitive distinctions during comprehension alone, in the attempt to inform our understanding of how the default encoding (i.e. the canonical meaning specification) of different quantifiers affects the recruitment of cognitive resources before any information relevant to verification is made available.

1.3 Current Study

This study is motivated by the belief that a better understanding of the default encoding of distinct quantifiers is essential if one wants to build a theory of how meaning is related to verification via cognitive resources. In fact, although the evidence for a link between representations of truth-conditions and verification is convincing, it is also evident that studying verification tasks alone can provide only *some* information about comprehension effects due to the encoding of distinct quantifiers. 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 & Zajenkowski (2011) suggest that monotonicity effects go in diverging directions with respect to comprehension and verification, depending on the cognitive task. Thus, we ask (a) whether there are effects of quantifier types on working memory during early comprehension, before subjects are allowed to engage in verification; and (b) whether they pattern as predicted by computational accounts of quantifier complexity.

Consistently with the main contrasts explored in previous studies, we selected quantifiers from four different categories (Aristotelian, Proportional, Numerical, Cardinal) according to their logical characterizations. We then evaluated the cognitive complexity of these 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 & Van Der Meer 2005; Karatekin *et al.* 2004; Ahern & Beatty 1979; Robison & Unsworth 2019). 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). Thus, pupillometry seems then to be a privileged technique to probe working memory demands as associated to the comprehension of quantified expressions.

Participants judged 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.

2 Methods

2.1 Apparatus

Eye movements and pupil area were recorded using an SR Research EyeLink 1000 desktop system using 35 mm lens, at a sampling frequency of 500 Hz. 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; 1920×1080 resolution; 60 Hz refresh rate). Participants sat at a distance of approximately 90 cm from the screen in a room with a dim light setup, and used a chin rest to stabilize their head. The camera itself was 60 cm away from the eyes, so 30 cm 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 or inaccurate eye-tracking calibration. Accuracy for the whole task was expected to reach a minimum of at least 85%. All participants fulfilled this criterion. Thus, 17 participants (male:3; female:14) were included in the final analyses.

2.3 Procedure

The experiment consisted of a short practice session (4 trials) followed by four experimental blocks. Each block was balanced so to contain approximately the same

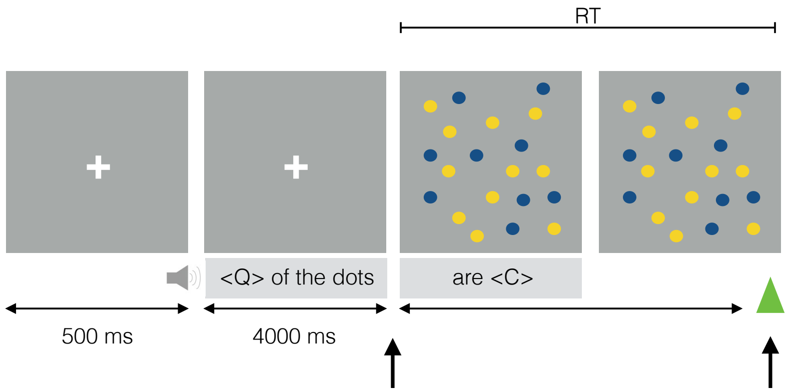


Figure 1: Experimental design.

number of trials for each quantifier. At the beginning of each block, 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 500 ms 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 4000 ms 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.* (2018)) before subjects could engage in verification. The onset of the disambiguating predicate was timed to 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 20 ms 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 — both the order of each block and the internal order of items within block were randomized across participants.

Participants were asked to express their judgment about the truth-value of the sentence by pressing a key (*f* or *j* — false and true, respectively) after the presentation of display. Participants were instructed to react as quickly as possible, but no time constraint was given for the decision phase, and the visual display stayed on until a decision was reached. The average length of the whole task was 1 hour. The experimental design and the time course of individual trials are shown in Figure 1.

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

Quantifier	Magnitude	Quantifier Category
All No Some		Aristotelian
At least n At most n	$n = 2, \ldots, 7; 9 \ldots 14$ $n = 2, \ldots, 7; 9 \ldots 14$	Numerical
An even number of An odd number of		Parity
Most More than half		Proportional

Table 1: Quantifiers grouped by category

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 14 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. To control for gaze shifts, the visual array was centered with respect to the gray background. 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, from a male native speaker of American English in his mid 20s.

2.5 Data analysis

SR Research DataViewer was used to output trial reports for three distinct interest periods: baseline (0-500 ms), encoding (500-4500 ms), and verification (4500 ms to key-press).[1] 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 out 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.* 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 Category (4 levels) and Proportion (14 levels) as fixed effects, and Participant as a random effect.

[1] Stimuli and raw data are available at https://github.com/aniellodesanto/PupillometryQuantifiers.

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

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. 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 Fig. 2), with no significant differences between RTs associated to Parity and Numerical quantifiers ($p < 0.986$).

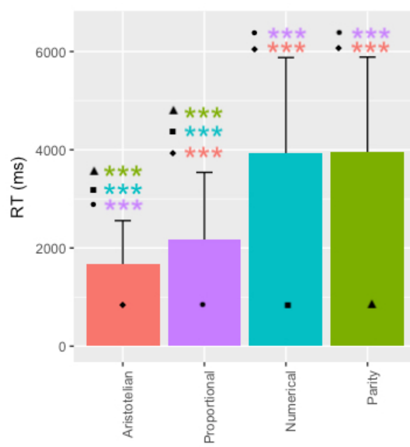
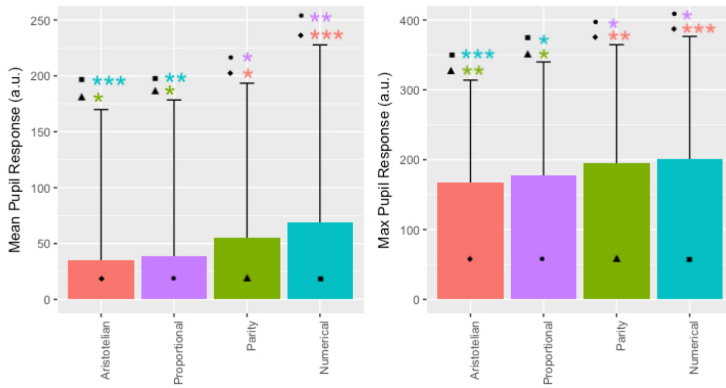


Figure 2: Comparisons of means by quantifier category for RT (in milliseconds) during verification. Signif. codes ($*** : 0.001; ** : 0.01; * : 0.05$) are coded based on the quantifier category of reference.

3.2 Pupillometry Results

3.2.1 Encoding

The linear mixed effects model and subsequent analysis of variance revealed significant effects of Quantifier Category 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: $F(15, 3190) =$



(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 coded based on the quantifier category of reference.

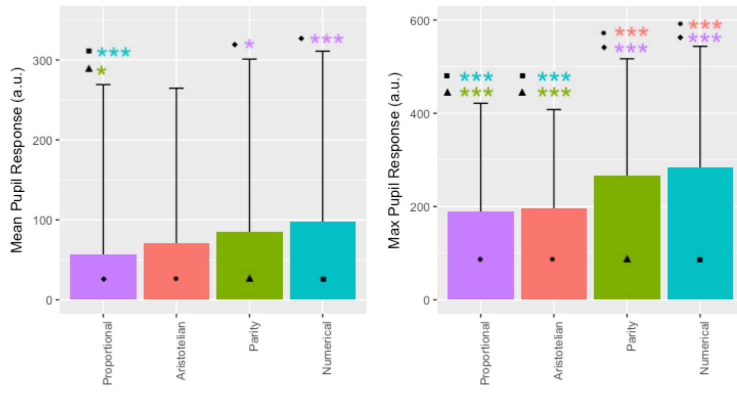
$0.86, p < 0.611$; max: $F(15, 3190) = 0.62, p < 0.858$). Post hoc Tukey comparison of means showed that quantifier effects cluster in two main groups, with Aristotelian and Proportional quantifiers eliciting significantly smaller pupil responses than Parity and Numerical ones (see Figure 3). No significant 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.

3.2.2 Verification

Significant effects were found of Quantifier Category 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.

4 Discussion

This paper presents an exploratory pilot study, employing 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 kinds of quantifiers (namely, Aristotelian, Proportional, Numerical, and Parity) could be found during early *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.



(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 coded based on the quantifier category of reference.

With respect to our main question, significant effects of Quantifier Category on pupil response during the encoding period support the hypothesis that working memory is in fact being modulated by quantifier meaning even before participants could engage in any type of verification strategy. While our small data sample advices caution in the interpretation of these results, we believe that the paradigm we employed highlights insightful patterns. A careful analysis of these effects can then shed light on the default encoding of generalized quantifiers, and how it is related to the recruitment of cognitive resources during verification.

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 initially 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). 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 (Heim *et al.* 2016:a.o.). 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 (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 not surprising then 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. Overall then, these effects support the hypothesis that the initial specification of Aristotelian and Proportional quantifiers recruits resources consistent with algorithms grounded in numerical estimation (possibly consistently with the assumptions of the ANS). Obviously, the absence of a difference between the Aristotelian and Proportional conditions could simply be due to our low sample size, and we should be careful in overextending the interpretation of these results. Much work needs to be done to fully determine whether Proportional quantifiers are represented in the same way as Aristotelian quantifiers, or if there is some intermediate level of representation.

Finally, similar response patterns for mean pupil response and for max response peak are found during the verification phase. RTs are also in line with this pattern: Aristotelian quantifiers are associated to the shortest RTs, and Numerical/Parity quantifiers to the longest ones. Together with the fact that pupil variation was still not significantly affected by the proportions of target colors, these results suggest that how the verification procedure is carried out for distinct quantifiers plays a less crucial role in modulating cognitive load than previously reported.

Recall now that the complexity hierarchy proposed by the semantic automata model sees Proportional quantifiers recruiting significantly more resources than Numerical and Parity, and it is thus in contrast to the results presented in this paper. In this sense, the pattern of RTs is particularly interesting, since it seems to contradict previous studies reporting processing effects mirroring this hierarchy.

A few considerations have to be made in this regard. First, as mentioned before, 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 (Emrich *et al.* 2009:a.o.). This suggests that RTs collected at the end of a decision task 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).

In this perspective, it is interesting to give a more careful look to our own results. While RTs for Proportional quantifiers overall pattern similarly to pupil responses — and are significantly shorter than those for Numerical/Parity quantifiers — they also show significant differences with Aristotelian quantifiers. This apparent mismatch between pupil response and RTs is consistent with the idea 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 procedures. 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.

These considerations also suggest that, 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). Then, we would 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.

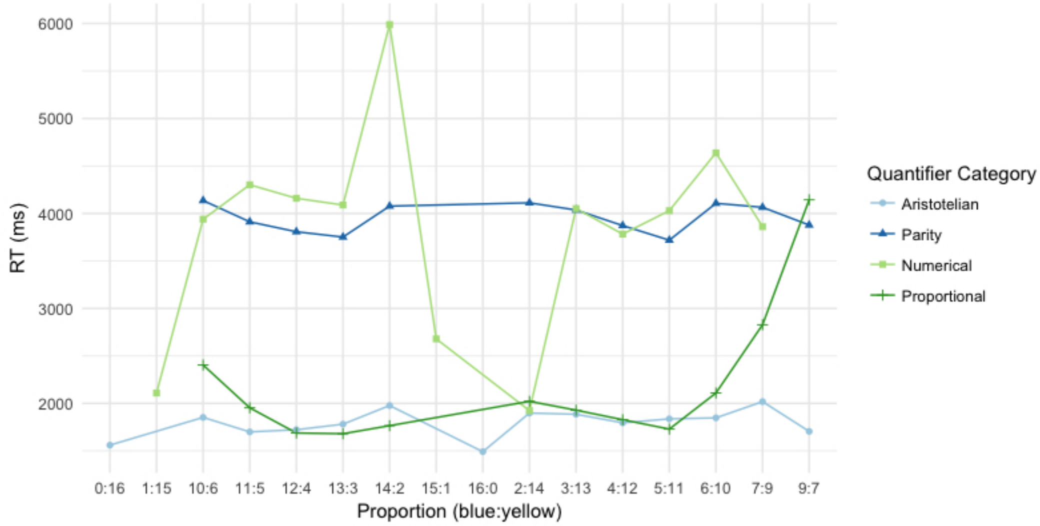


Figure 5: Comparisons of RT by quantifier category (in milliseconds) and proportions of colored dots (*blue:yellow*) during verification.

Relatedly, the semantic automata's predictions have been observed to hold when approximate numerical estimation is explicitly disfavored by the verification context (e.g., the visual scene). In contrast, we compared proportional quantifiers over a range of varying proportions. Furthermore, 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. On the other hand, when precise counting is discouraged across quantifiers, we predict a replication of this paper's results. These hypotheses should be better investigated in future studies, with particular focus on probing eventual differences between cognitive load as measured by RTs and pupil response — for instance, by following the design of Heim *et al.* (2012). Similarly, this study cannot fully rule out potential non-semantic, low-level contributions to the difference between quantifier categories (e.g., syllabic length of a quantified sentence). These limitations will have to be carefully addressed in future work.

References

Ahern, S., & J. Beatty. 1979. Pupillary responses during information processing vary with scholastic aptitude test scores. *Science* 205.1289–1292.

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

Barwise, J., & R. Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy* 4.

Clark, R., & M. Grossman. 2007. Number sense and quantifier interpretation. *Topoi* 26.51–62.

Degen, J., & M. K. Tanenhaus. 2016. Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science* 40.

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

Dotlacil, J., J. Szymanik, & M. Zajenkowski. 2014. Probabilistic semantic automata in the verification of quantified statements. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society, July 23-26, 2014*.

Emrich, S. M., N. Al-Aidroos, J. Pratt, & S. Ferber. 2009. Visual search elicits the electrophysiological marker of visual working memory. *PLOS ONE* 4.

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

Feigenson, L., S. Dehaene, & E. Spelke. 2004. Core systems of number. *Trends in cognitive sciences* 8.307–314.

Halberda, J., & L. Feigenson. 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., K. Amunts, D. Drai, S. Eickhoff, S. Hautvast, & Y. Grodzinsky. 2012. The language number interface in the brain: A complex parametric study of quantifiers and quantities. *Frontiers in Evolutionary Neuroscience* 4.

Heim, S., C. T. McMillan, R. Clark, L. Baehr, K. Ternes, C. Olm, N. E. Min, & M. Grossman. 2016. How the brain learns how few are “many": An fMRI study of the flexibility of quantifier semantics. *NeuroImage* 125.

Just, M. A., P. A. Carpenter, & A. Miyake. 2003. Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related potential studies of brain work. In *Theoretical Issues in Ergonomics Science*.

Karatekin, C., J. W. Couperus, & D. J. Marcus. 2004. Attention allocation in the dual-task paradigm as measured through behavioral and psychophysiological responses. *Psychophysiology* 41.175–185.

Laeng, B., S. Sirois, & G. Gredebäck. 2012. Pupillometry: a window to the preconscious? *Perspectives on psychological science* 7.18–27.

Lidz, J., P. Pietroski, J. Halberda, & T. Hunter. 2011. Interface transparency and the psychosemantics of most. *Natural Language Semantics* 19.

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

Nuthmann, A., & E. Van Der Meer. 2005. Time’s arrow and pupillary response. *Psychophysiology* 42.306–317.

Paterson, K. B., R. Filik, & L. M. Moxey. 2009. Quantifiers and discourse processing. *Language and Linguistics Compass* 3.1390–1402.

Pietroski, P. M. 2010. Concepts, meanings and truth: First nature, second nature and hard work. *Mind & Language* .

Pietroski, P. M., J. Lidz, T. Hunter, & J. Halberda. 2009. The meaning of “most": Semantics, numerosity and psychology. *Mind & Language* 24.

Robison, M. K., & N. Unsworth. 2019. Pupillometry tracks fluctuations in working memory performance. *Attention, Perception, & Psychophysics* 81.

Shikhare, S., S. Heim, E. Klein, S. Huber, & K. Willmes. 2015. Processing of numerical and proportional quantifiers. *Cognitive Science* 39.1504–1536.

Stanners, R. F., M. Coulter, A. W. Sweet, & P. Murphy. 1979. The pupillary response as an indicator of arousal and cognition. *Motivation and Emotion* 3.319–340.

Steinert-Threlkeld, S., G.-J. Munneke, & J. Szymanik. 2015. Alternative representations in formal semantics: A case study of quantifiers. In *Proceedings of the 20th Amsterdam Colloquium*, ed. by N. T. T. Brochhagen, F. Roelofsen.

Szymanik, J., & M. Zajenkowski. 2010. Comprehension of simple quantifiers: Empirical evaluation of a computational model. *Cognitive Science* 34.521–532.

Szymanik, J., & M. Zajenkowski. 2011. Contribution of working memory in parity and proportional judgments. *Belgian Journal of Linguistics* 25.176–194.

Troiani, V., J. Peelle, R. Clark, & M. Grossman. 2009. Is it logical to count on quantifiers?Dissociable neural networks underlying numerical and logical quantifiers. *Neuropsychologia* .

Van Benthem, J. 1986. *Quantifiers*. Dordrecht: Springer Netherlands.

Zajenkowski, M., R. Styła, & J. Szymanik. 2011. A computational approach to quantifiers as an explanation for some language impairments in schizophrenia. *Journal of Communication Disorders* 44.595–600.

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