Encoding and Verification Effects of Generalized Quantifiers on Working Memory

Aniello De Santo (aniello.desanto@stonybrook.edu)
Department of Linguistics
Stony Brook University

John E. Drury (john.drury@qu.edu.qa)
Department of English Literature and Linguistics
Qatar University.

Abstract

A large amount of literature has shown that the type of quantifiers 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.

Keywords: Generalized Quantifiers; Default Encoding; Working Memory; ANS; Semantic Automata

Introduction

Barwise and 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, c.f. (Lidz, Pietroski, Halberda, & Hunter, 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 (Degen & Tanenhaus, 2016; Pietroski, 2010; Steinert-Threlkeld, Munneke, & Szymanik, 2015, a.o.).

One model in numerical cognition that has significantly influenced current theories of quantifiers' interpretation is grounded in the Approximate Number System (ANS) — suggested to explain the representation of imprecise cardinalities, and thus the ability to compare quantities without counting.

Past research has explored the idea that quantifier comprehension can be conceptualized with the aid of numerical comparison rooted in the ANS (Dehaene, 1999). In particular, 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. 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 (Heim et al., 2012, 2016). Additionally, recent studies have noticed variability among verification strategies associated to different quantifiers which, in contexts in which numerical estimation is disfavored by the visual scene, seems to be consistent with the predictions made by verification algorithms grounded in precise counting strategies (Szymanik, 2016, a.o.).

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. Here, to specifically disentangle early comprehension (encoding) from verification, we used pupillometry measures to probe the effect of processing the meaning of distinct classes of quantifiers on working memory resources.

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, Szymanik, & Zajenkowski, 2014). Thus, we ask whether there are any any effects of quantifier types on working memory during early comprehension, before subjects are allowed to engage in verification.

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 (Robison & Unsworth, 2019, a.o.). Variations in pupil size have also been widely used as an estimate of working memory in visual search tasks (Just, Carpenter, & Miyake, 2003), and have been shown to be sensitive to local resource demands imposed by sentence comprehension (Engelhardt, Ferreira, & Patsenko, 2010). Thus, pupillometry seems then to be a privileged technique to probe working memory demands as associated to the comprehension of quantified expressions.

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.

Methods

Apparatus

Eye movements and pupil area were recorded using an SR Research EyeLink 1000 desktop system using 35mm lens, at a sampling frequency of 500 Hzszymanik2016book. 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). All viewers 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.

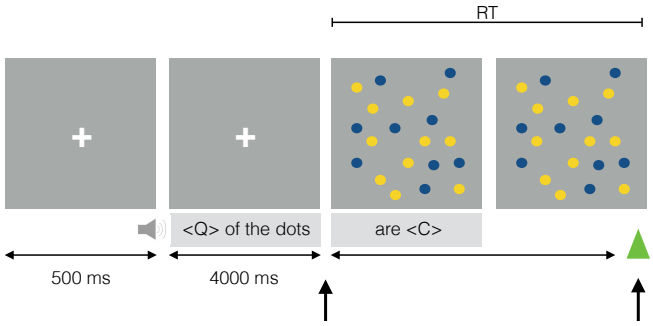


Figure 1: Experimental design.

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 participant fulfilled this criterion. Thus, 17 participants (male:3; female:14) were included in the final analyses.

Procedure

The experimental design we used to address these questions is shown in Figure 1. The experiment consisted of a short practice session (4 trials) followed by four experimental blocks. 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, Fabius, Van Heusden, & Van der Stigchel, 2018)) before subjects could engage in verification. The onset of the disambiguating predicate 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 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.

Participants were asked to express their judgment about the truth-value of the sentence by pressing a key (*f* or *j*, associ-

Table 1: Quantifiers grouped by category

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

ated with false and true respectively) at any point 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.

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.

Data analysis[1]

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). 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., 2018; Attar, Schneps, & Pomplun, 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 to be shared on the authors' website.

Table 2: Accuracy Results

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

Results

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

Pupillometry Results

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: $F(15,3190) = 0.86$, $p < 0.611$; max:

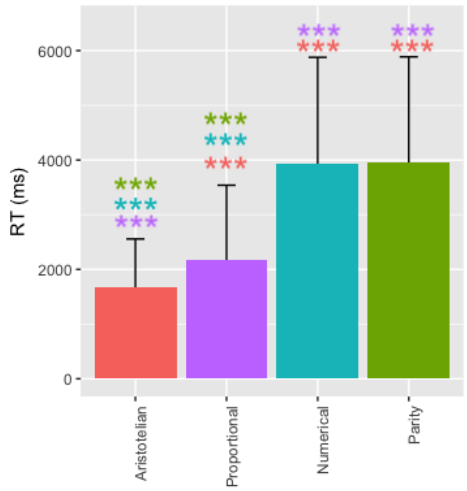


Figure 2: Comparisons of means by quantifier category for RT (in milliseconds) during verification. Signif. codes ($***$: 0.001; $**$: 0.01; $*$: 0.05) are color coded based on the quantifier category of reference.

$F(15,3190) = 0.62$, $p < 0.858$). 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.

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.

Discussion

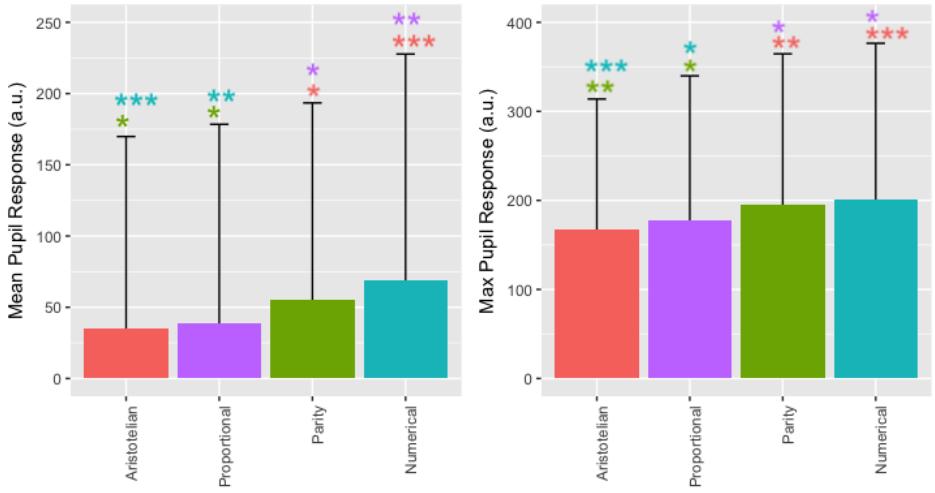
This exploratory study employed recordings of pupil size variation during a truth-value judgment task to better understand cognitive resources underlying the processing of quantified sentences. In particular, we were interested in exploring whether effects of different kind of quantifiers (namely, Aristotelian, Proportional, Numerical, and Parity) could be found during 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.

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. A careful analysis of these effects can then shed light on the default encoding of generalized quantifiers belonging to different classes, and how it is related to the recruitment of cognitive resources by different verification algorithms.

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, Peelle, Clark, & Grossman, 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, Lidz, Hunter, & Halberda, 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 is grounded in the ANS, and recruits resources consistent with algorithms grounded in numerical estimation.

Similar response patterns both for mean pupil response and for max response peak are found during the verification phase. RTs also showed a similar pattern: with Aristotelian quantifiers being associated to the shortest RTs, and Numerical/Parity quantifiers to the longest ones. These results, 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.

However, the pattern of RTs is particularly interesting, since it seems to contradict a set of studies on quantifier verification, which reported processing effects mirroring the complexity hierarchy proposed by models based on precise counting strategies. In particular, Szymanik (2016) provides a significant amount of evidence supporting the semantic au-



(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.

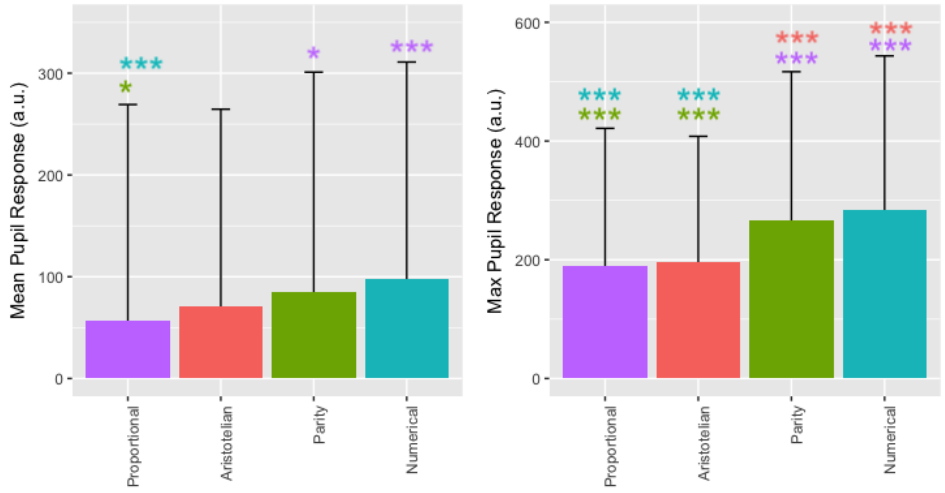
tomata framework as a good model of how cognitive resources are engaged by different quantifiers, based on a verification task over a visual scene. This model associates quantifiers to computational mechanisms (automata) implementing specific recognition procedures employed for the verification process, via an algorithmic approach based on counting (Van Benthem, 1986), and then places quantifiers in a hierarchy based on the complexity of the machines required for their verification. Crucially, this hierarchy sees Proportional quantifiers recruiting significantly more resources than Numerical and Parity, and it is thus in contrast to the results presented in this paper.

A few considerations have to be made in this regard. First of all, the semantic automata predictions have been observed to hold when approximate numerical estimation is explicitly disfavored by the verification context (e.g. the visual scene). In our experiment, the proportion of the target set ranged from cardinalities close to each other to cardinalities far apart, in order to avoid possible biases for one strategy over the other. In fact, this seems to add strength to our results as evidence that the default encoding of Proportional quantifiers relies on approximation strategies. In the future though, it would be interesting to test the effect of varying proportions on pupil response and RTs more systematically, possibly following the design of (Heim et al., 2012).

Secondly, 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 (Emrich, Al-Aidroos, Pratt, & Ferber, 2009, a.o.) and suggest 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. We believe this apparent mismatch between pupil response and RTs suggests 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. These hypotheses should be better investigated in future studies, with particular focus on probing eventual differences between cognitive load



(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.

as measured by RTs and pupil response.

References

Attar, N., Schneps, M. H., & Pomplun, M. (2016). Working memory load predicts visual search efficiency: Evidence from a novel pupillary response paradigm. *Memory & Cognition*(DOI: 10.3758/s13421-016-0617-8).

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*(2).

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, *40*(1). doi: 10.1111/cogs.12227

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

Dotlacil, J., Szymanik, J., & Zajenkowski, M. (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., Al-Aidroos, N., Pratt, J., & Ferber, S. (2009, 11). Visual search elicits the electrophysiological marker of visual working memory. *PLOS ONE*, *4*(11).

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology*, *63*(4).

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*. doi: 10.3389/fnevo.2012.00004

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

Just, M. A., Carpenter, P. A., & Miyake, A. (2003). Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related potential studies of brain work. In *Theoretical issues in ergonomics science.*

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of most. *Natural Language Semantics*, *19*(3).

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*.

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

Pietroski, P. M., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of "most": Semantics, numerosity and psychology. *Mind & Language*, *24*(5).

Robison, M. K., & Unsworth, N. (2019, 2). Pupillometry tracks fluctuations in working memory performance. *Attention, Perception, & Psychophysics*, *81*(2).

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

Szymanik, J. (2016). *Quantifiers and cognition: Logical and computational perspectives* (Vol. 96). Springer.

Troiani, V., Peelle, J. E., Clark, R., & Grossman, M. (2009). Is it logical to count on quantifiers? Dissociable neural networks underlying numerical and logical quantifiers. *Neuropsychologia*, *47*(1).

Van Benthem, J. (1986). Quantifiers. In *Essays in logical semantics.* Dordrecht: Springer Netherlands.