Cognition 223 (2022) 105013

Contents lists available at ScienceDirect

Cognition

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

Computational complexity explains neural differences in quantifier verification

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

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

ARTICLE INFO

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

ABSTRACT

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

1. Introduction

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

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

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

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

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

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

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

1.1. Quantifier automata and the computational complexity of verification

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

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

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

1.2. Previous research and relevant electrophysiological effects

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

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

Additionally, few studies have looked at sentence verification in relation to a picture. Spychalska et al. (2019, 2016) were only interested in sentence final effects of implicature violations, and showed the picture mid-sentence, immediately before the final word. This modulated the N400 and post-N400 positivities. The authors were able to show that participants' pragmatic sensitivity had an effect on the evoked potential in trials where scalar implicatures were modulated. However, the design did not allow investigating incremental effects of verification that could originate at earlier points in the sentence. Hunt III et al. (2013) and Politzer-Ahles et al. (2013) were also interested in implicature violations, but presented pictures before each sentence. The former found graded N400 responses with a visual world paradigm for true, underinformative and false sentences: false sentences elicited the strongest effect compared to true, whereas underinformative fell in the middle. Politzer-Ahles et al. (2013) looked at effects on the quantifier. In a 2×2 design with 'some' and 'all' – where 'all' was true when 'some' was

underinformative, and false when 'some' was strictly true – they found sustained positivities for quantificational violations with 'all', but sustained negativities for implicature violations with 'some'. Augurzky et al. (2017, 2019, 2020a, 2020b) have all addressed issues of incrementality. They found that, regardless of quantifier type – Aristotelian or proportional, in nominal, e.g., 'all the circles', or adverbial form, e.g., 'every day' – the N400, and related truth value effects, are only found at the position where the sentence is disambiguated. When the presented linguistic material is compatible with the sentence being both true and false, N400 effects do not arise. The only exception to this pattern is the negative proportional quantifier 'less than half', for which the N400 does not arise at all (see also Nieuwland, 2016; Urbach et al., 2015; Urbach & Kutas, 2010). In these cases, they instead found an increased positivity on the quantifier, which they attributed to the semantic complexity of the negative polarity (see e.g. Deschamps et al., 2015; Just & Carpenter, 1971). In all experiments, a sustained positivity was also found after the N400 in false trials where the truth value could not be known immediately, but only when participants performed a verification task. The authors attributed this to increased attention to the picture-sentence mapping in complex contexts, and argue that it is a P600-as-P3 decision effect (Sassenhagen et al., 2014).

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

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

1.3. The present study

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

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

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

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

2. Experiment 1

2.1. Method

2.1.1. Design

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

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

2.1.2. Participants

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

2.1.3. Materials

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

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

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

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

Table 1
Experiment sentences.

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

2.1.4. Procedure

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

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

2.1.5. EEG-recording

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

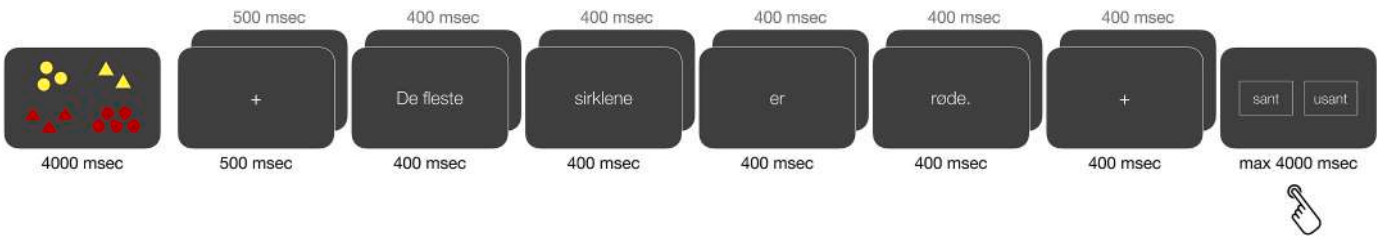

Fig. 1. Structure of a single trial from experiment 1. Trial structure was the same in experiment 2, except that the true/false (sann/usann) screen was replaced by a comprehension question (4000 ms) followed by a maximum 4000 ms interval within which the participant could produce an answer.

2.1.6. Data analysis

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

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

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

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

2.2. Results

2.2.1. Behavioral results

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

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

2.2.2. EEG results

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

Table 2
Accuracy and response times, Experiment 1.

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

Table 3
Logistic regression on accuracy, Experiment 1.

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

Table 4
Linear regression on response times, Experiment 1.

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

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

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

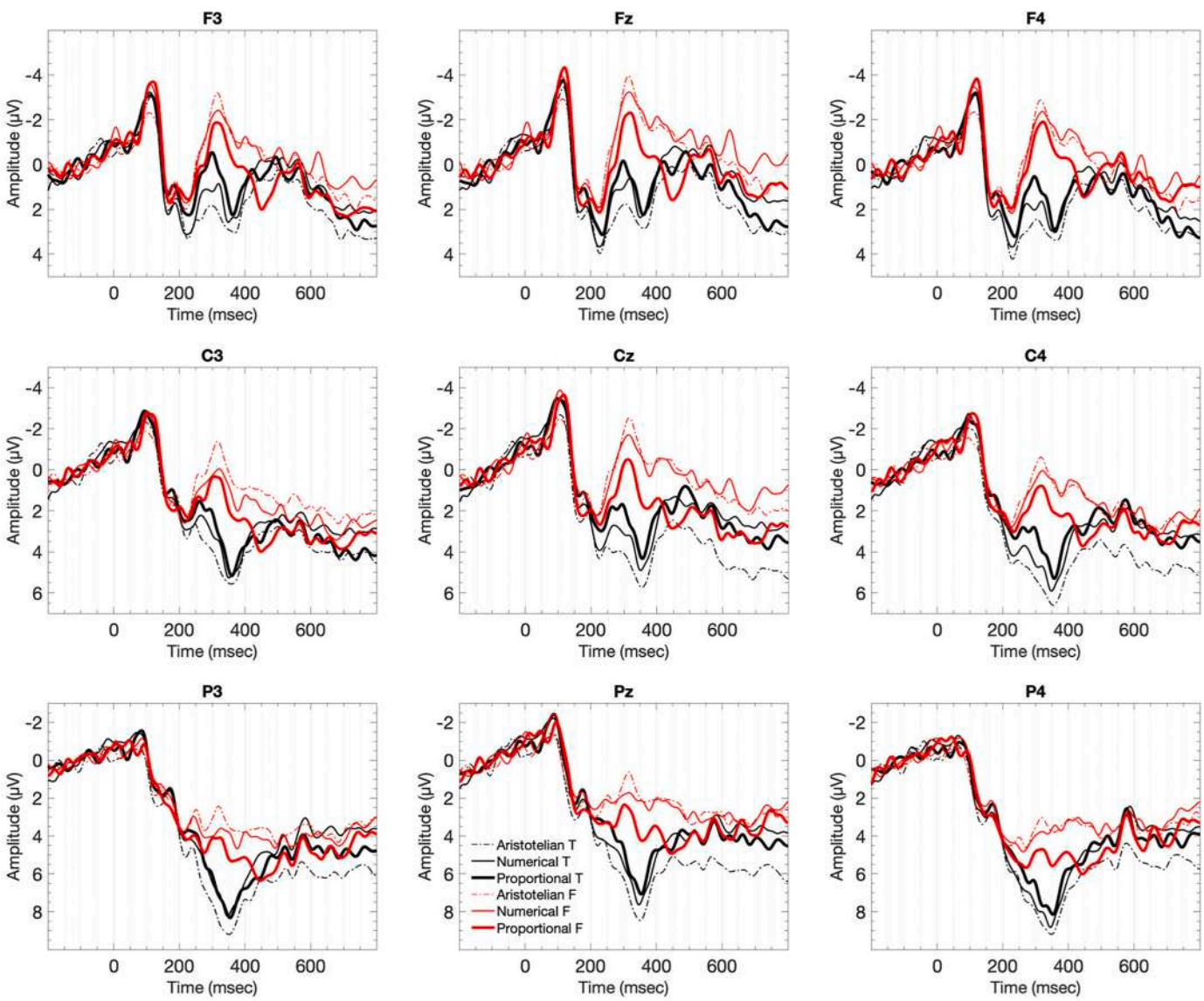

Fig. 2. Grand-average ERP waveforms from 9 selected channels, time locked to the onset of the sentence-final adjective (0 ms) in experiment 1. True trials are shown in black, false trials in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version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

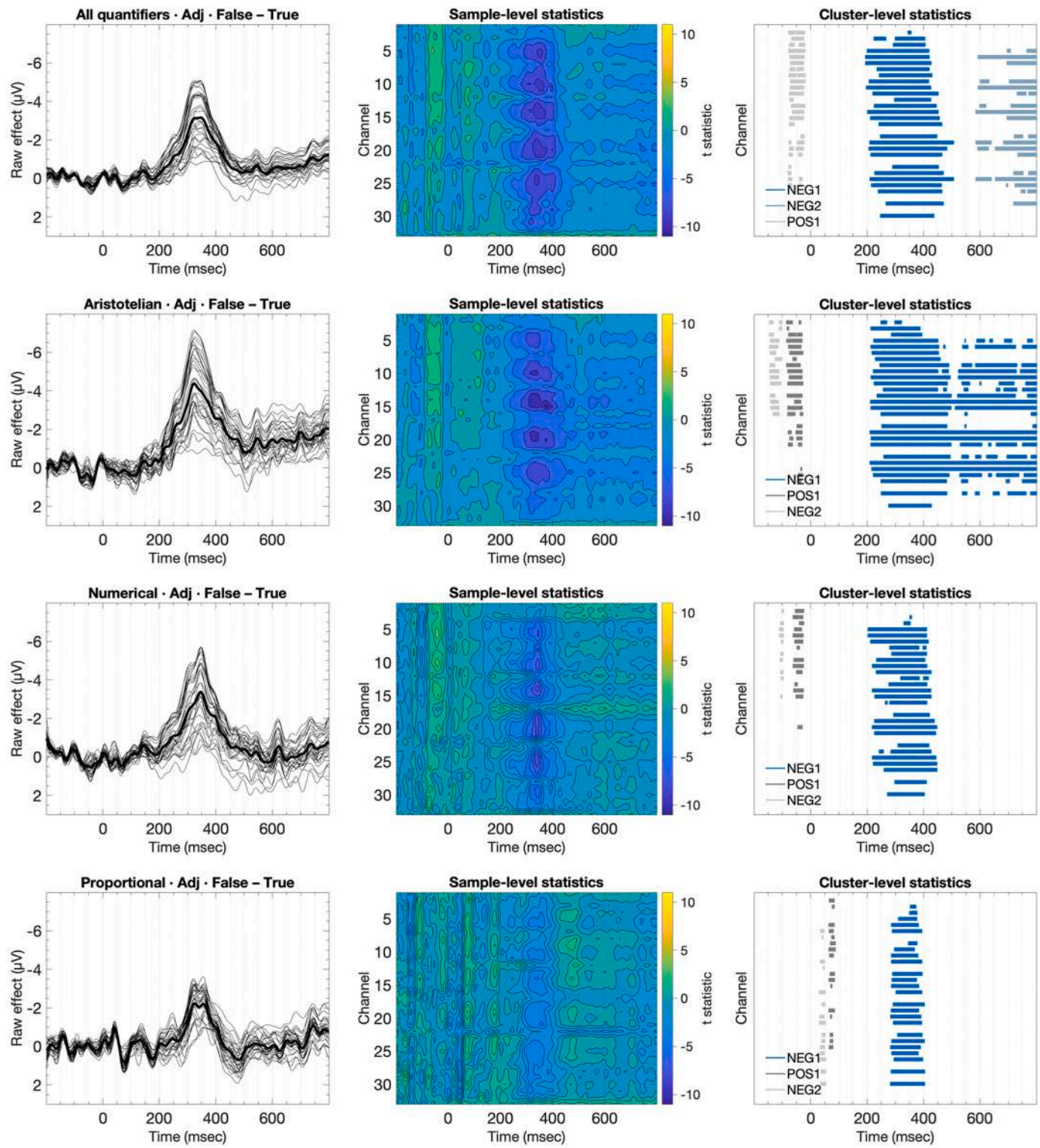

Fig. 3. ERP effects of truth value (False-True) across quantifier classes, time locked to the onset of the sentence-final adjective (0 ms) in experiment 1. Raw effect waveforms (left column) are displayed along with contour maps of sample-level statistics (middle column) and raster plots of cluster-level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

quantifier classes. All true trials present a clear P300 component, particularly visible over posterior channels (Fig. 2, black lines). The P300 component appears largest for true trials with Aristotelian quantifiers and smallest for true trials with proportional quantifiers, with numerical quantifiers falling in between. These differences persist throughout the epoch (Fig. 2). In direct comparisons between true trials across quantifier classes, we only found a marginal effect for the first-ranked cluster in the contrast between Aristotelian and proportional quantifiers ($T_{sum} = 2081.42$, $S = 806$, $p = 0.072$), and no effects for Aristotelian vs numerical or numerical vs proportional. These data

indicate that verification strategies at the sentence-final word for true trials do not differ, in terms of underlying physiology, between quantifier classes.

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

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

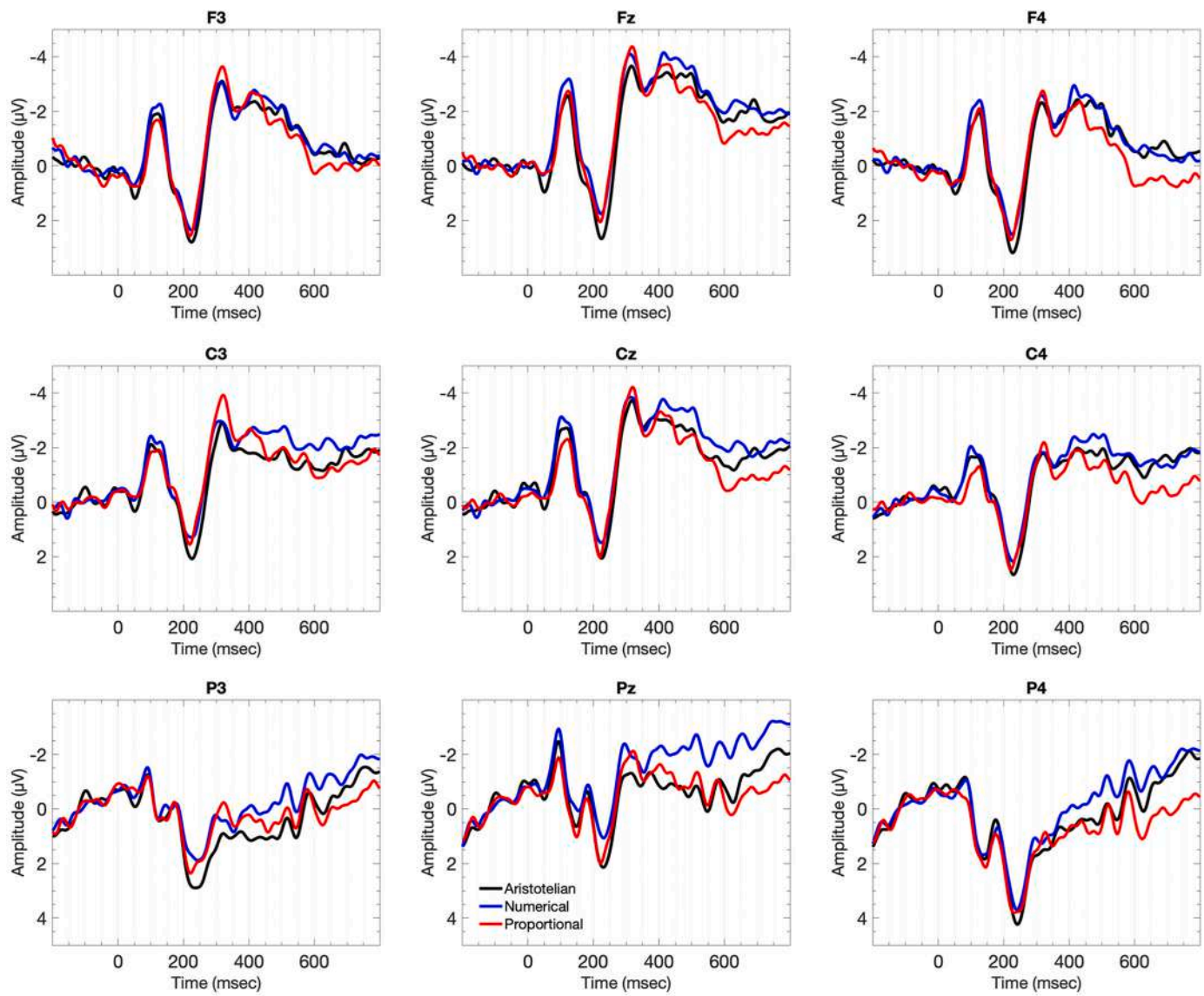

Fig. 4. Grand-average ERP waveforms from 9 selected channels, time locked to the onset of the sentence-internal noun (0 ms) in experiment 1. Trials from nouns following Aristotelian quantifiers are shown in black, blue is numerical quantifiers, and red is proportional quantifiers. (For interpretation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

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

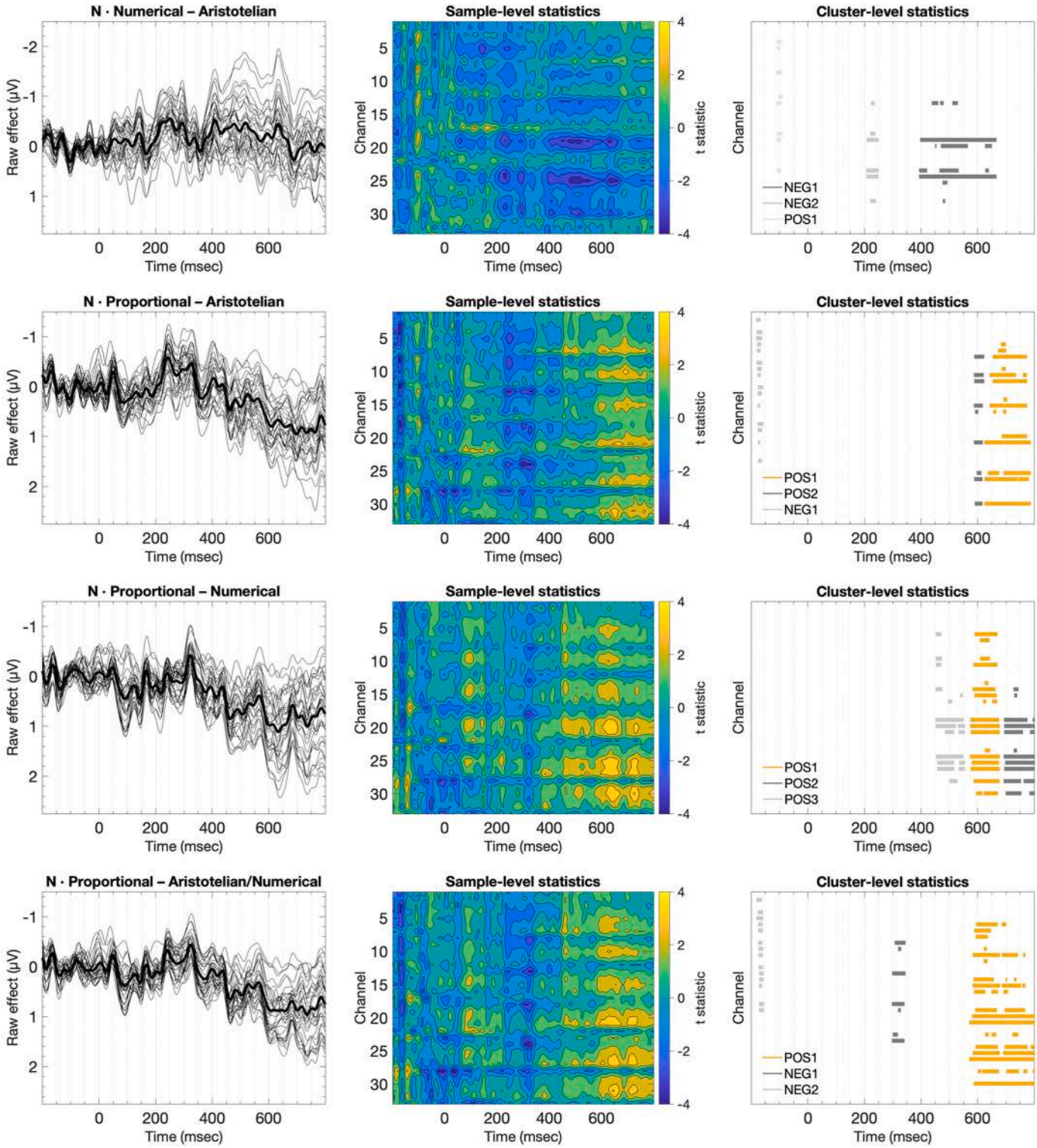

Fig. 5. ERP effects of pairwise comparisons between quantifier classes, time locked to the onset of the sentence-internal noun (0 ms) in experiment 1. Raw effect waveforms (left column) are displayed along with contour maps of sample-level statistics (middle column) and raster plots of cluster-level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

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

2.3. Interim discussion

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

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

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

3. Experiment 2

3.1. Method

3.1.1. Participants

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

3.1.2. Materials

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

3.1.3. Procedure

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

3.1.4. EEG-recording

There were no differences in EEG recording between experiments.

3.1.5. Data analysis

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

added as fixed effects.

3.2. Results

3.2.1. Behavioral results

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

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

3.2.2. EEG results

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

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

Table 5
Accuracy and response times, Experiment 2.

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

Table 6
Logistic regression on accuracy, Experiment 2.

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

Table 7
Linear regression on response times, Experiment 2.

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

3.3. Interim discussion

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

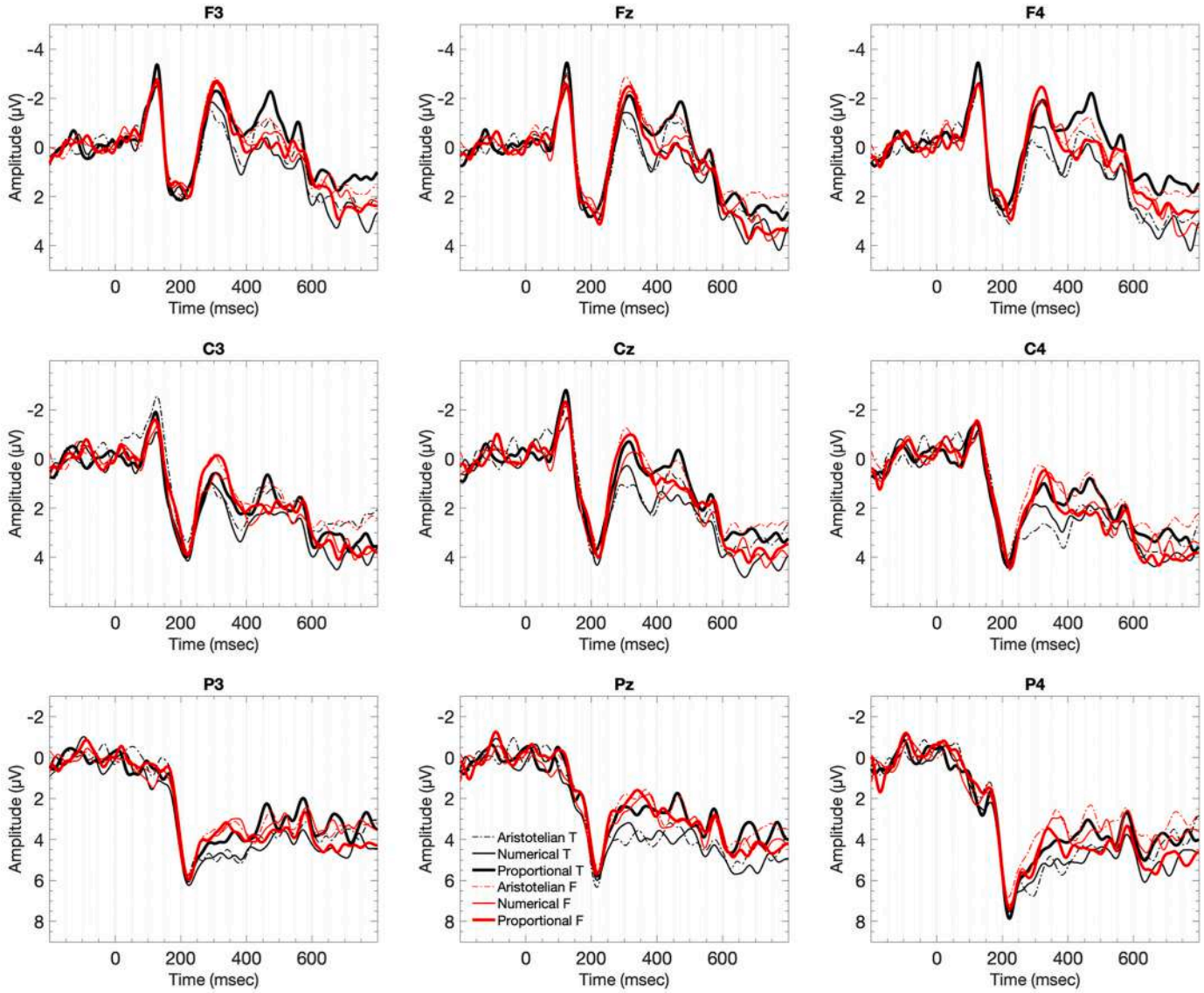

Fig. 6. Grand-average ERP waveforms from 9 selected channels, time locked to the onset of the sentence-final adjective (0 ms) in experiment 2. True trials are shown in black, false trials in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version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

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

4. General discussion

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

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

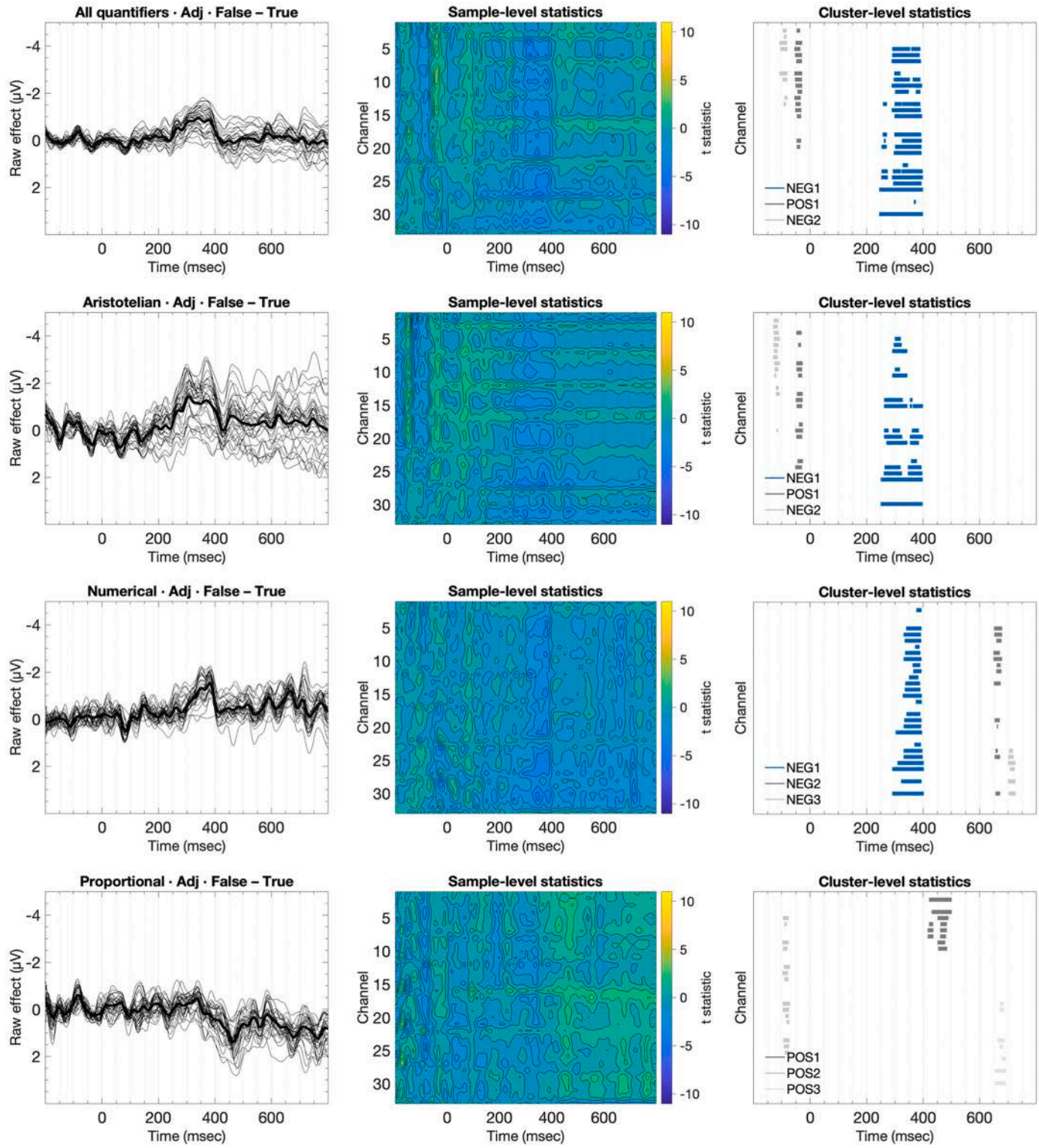

Fig. 7. ERP effects of truth value (False-True) across quantifier classes, time locked to the onset of the sentence-final adjective (0 ms) in experiment 2. Raw effect waveforms (left column) are displayed along with contour maps of sample-level statistics (middle column) and raster plots of cluster-level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

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

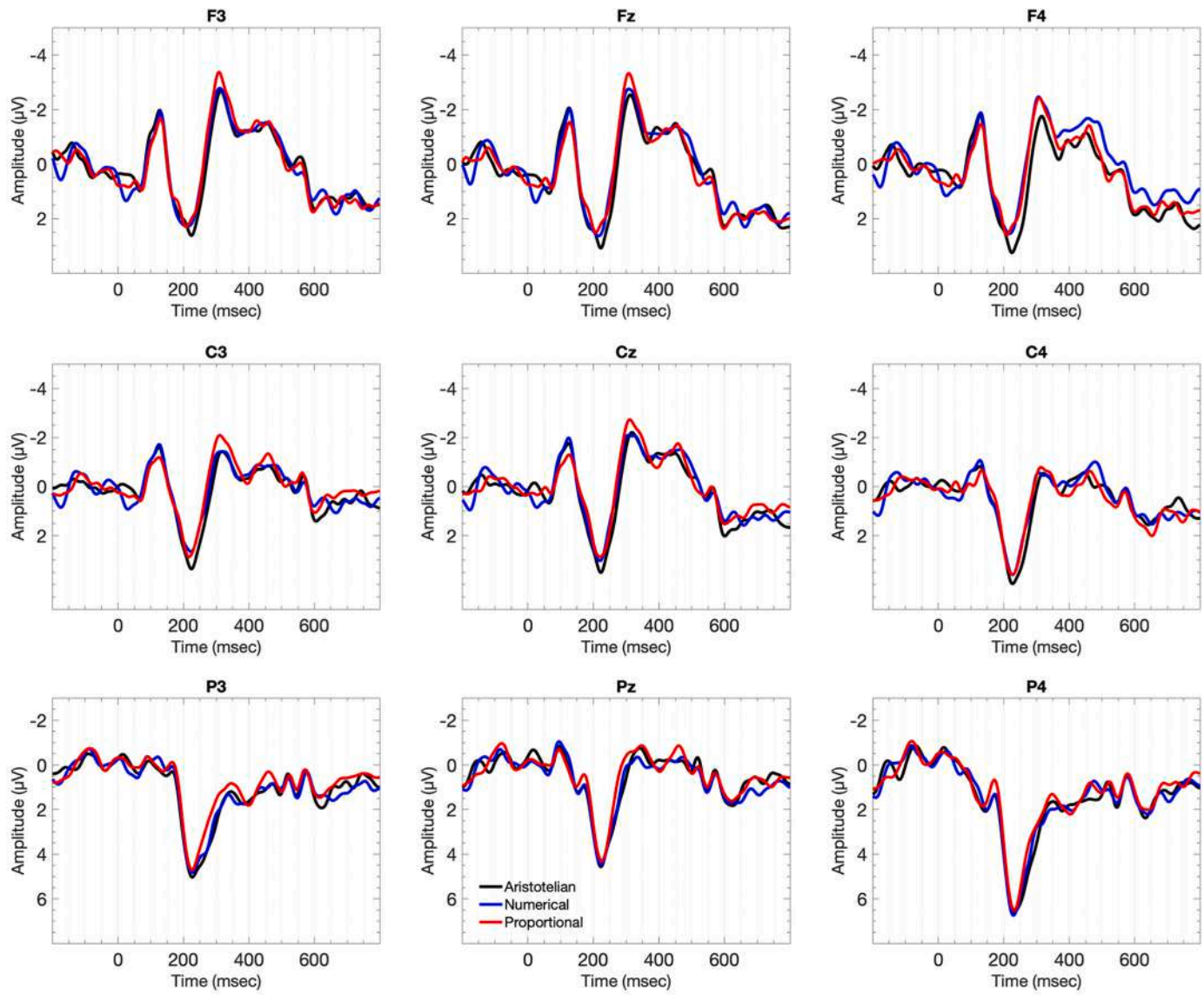

Fig. 8. Grand-average ERP waveforms from 9 selected channels, time locked to the onset of the sentence-internal noun (0 ms) in experiment 2. Trials from nouns following Aristotelian quantifiers are shown in black, blue is numerical quantifiers, and red is proportional quantifiers. (For interpretation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

remains to be explained.

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

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

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

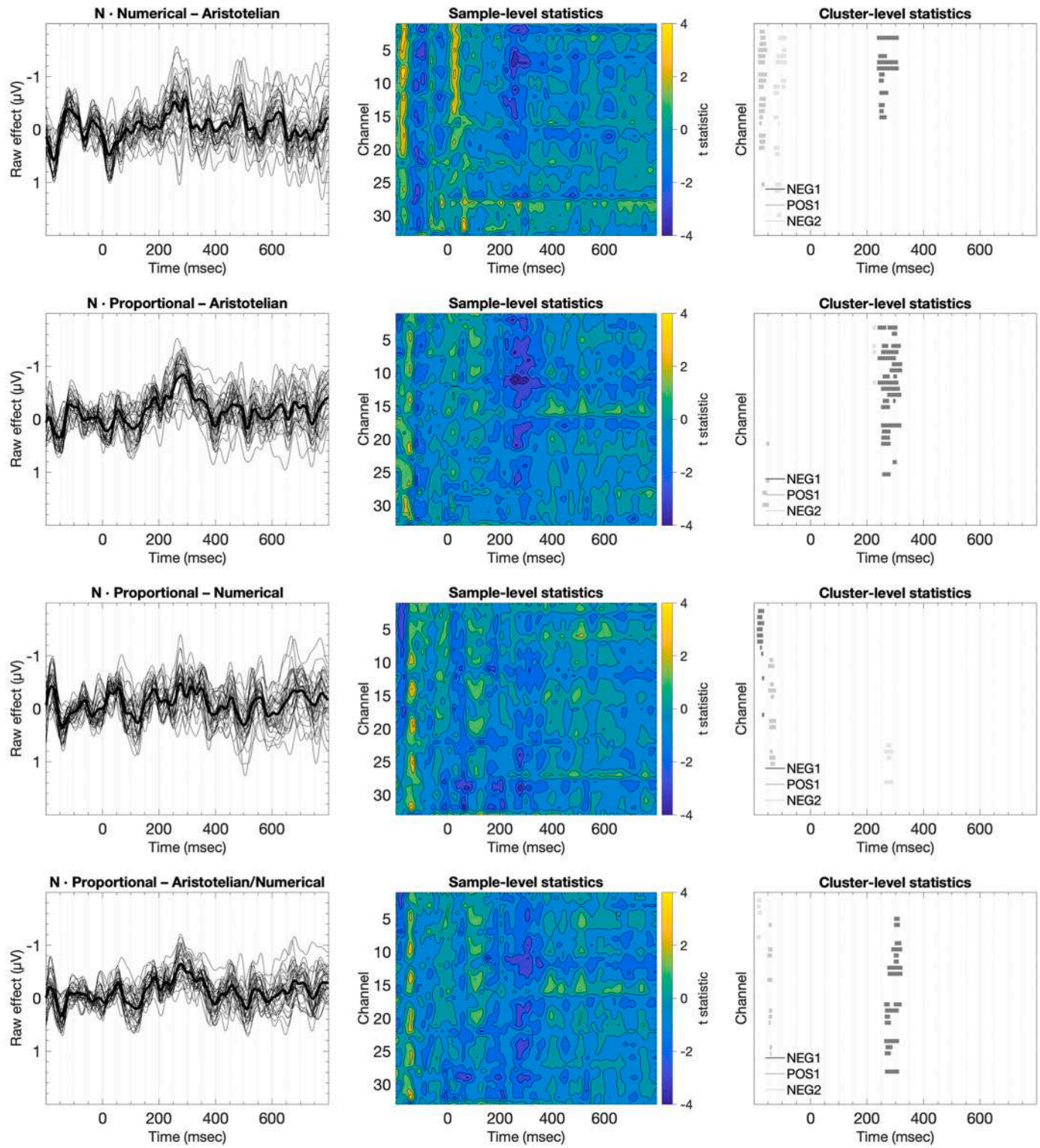

Fig. 9. ERP effects of pairwise comparisons between quantifier classes, time locked to the onset of the sentence-internal noun (0 ms) in experiment 2. Raw effect waveforms (left column) are displayed along with contour maps of sample-level statistics (middle column) and raster plots of cluster-level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

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

If a model of sentence meaning has been built at the final word, then the positivity at the noun can be argued to be a signature of verification. The time-course and distribution of the effect is similar to the LPC

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

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

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

5. Conclusion

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

Data Availability Statement

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

Acknowledgements

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

Appendix A. Supplementary data

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

References

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

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

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

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

Bach, E., Jelinek, E., Kratzer, A., & Partee, B. H. (1995). *Quantification in natural languages*. Dordrecht: Kluwer Academic Publishers. editors.

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

Baggio, G. (2020). Epistemic transfer between linguistics and neuroscience: Problems and prospects. In R. Nefdt, C. Klippi, & B. Karstens (Eds.), *The philosophy and science of language: Interdisciplinary perspectives, pages* (pp. 275–308). Cham: Palgrave Macmillan.

Baggio, G. (2021). Compositionality in a parallel architecture for language processing. *Cognitive Science, 45*(5), e12949.

Baggio, G., & Bremnes, H. S. (2017). Book review: Jakub Szymanik quantifiers and cognition. Logical and computational perspectives. Springer, 2016. *Studia Logica, 105*, 1015–1019.

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

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

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bremnes, H. S. (2021). *Data for computational complexity explains neural differences in quantifier verification*. DataverseNO. https://doi.org/10.18710/M6VT6Z.

Brodbeck, C., Gwilliams, L., & Pylkkänen, L. (2016). Language in context: MEG evidence for modality-general and -specific responses to reference resolution. *eNeuro, 3*.

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

Carcassi, F., Steinert-Threlkeld, S., & Szymanik, Jakub (2021). Monotone quantifiers emerge via iterated learning. *Cognitive Science, 45*(8).

Chemla, E., Dautriche, I., Buccola, B., & El Fagot, J. (2019). Constraints on the lexicons of human languages have cognitive roots present in baboons (*Papio papio*). *PNAS, 116*, 30.

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

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

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*, 472–517.

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

Coppock, E. (2019). Quantity superlatives in Germanic, or life on the fault line between adjective and determiner. *Journal of Germanic Linguistics, 31*, 109–200.

D'Arcy, R. C. N., Connolly, J. F., & Crocker, S. F. (2000). Latency shifts in the N2b component track phonological deviations inspoken words. *Clinical Neurophysiology, 111*, 40–44.

De Santo, A., Rawski, J., Yazdani, A. M., & Drury, J. E. (2019). Quantified sentences as a window into prediction and priming: An ERP study. In E. Ronai, L. Stigliano, & Y. Sun (Eds.), *Proceedings of the fifty–fourth annual meeting of the Chicago linguistic society* (pp. 85–98).

Deschamps, I., Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition, 143*, 244–253.

Dwivedi, V. D., Phillips, N. A., Einagel, S., & Baum, S. R. (2010). The neural underpinnings of semantic ambiguity and anaphora. *Brain Research, 1311*, 93–109.

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

Faarlund, J. T., Lie, S., & Vannebo, K. I. (1997). *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

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

Fritz, I., & Baggio, G. (2020). Meaning composition in minimal phrasal contexts: Distinct ERP effects of intensionality and denotation. *Language, Cognition and Neuroscience, 35*(10), 1295–1313.

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

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

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

Hubbard, R. J., Rommers, J., Jacobs, C. L., & Federmeier, K. D. (2019). Downstream behavioral and electrophysiological consequences of word prediction on recognition memory. *Frontiers in Human Neuroscience, 13*, 291.

Hunt, L., III, Politzer-Ahles, S., Gibson, L., Minai, U., & Fiorentino, R. (2013). Pragmatic inferences modulate N400 during sentence comprehension: Evidence from picture-sentence verification. *Neuroscience Letters, 534*, 246–251.

Hunter, T., & Lidz, J. (2013). Conservativity and learnability of determiners. *Journal of Semantics, 30*, 315–334.

Hunter, T., Lidz, J., Odic, D., & Wellwood, A. (2017). On how verification tasks are related to verification procedures: A reply to Kotek et al. *Natural Language Semantics, 25*, 91–107.

Isaac, A. M. C., Szymanik, J., & Verbrugge, R. (2014). Logic and complexity in cognitive science. In A. Baltag, & S. Smets (Eds.), *Johan Van Benthem on logic and information dynamics* (pp. 787–824).

Jacob, S. N., & Nieder, A. (2009). Notation-independent representation of fractions in the human parietal cortex. *Journal of Neuroscience, 29*(14), 4652–4657.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge: Cambridge University Press.

Just, M. A. (1974). Comprehending quantified sentences: The relation between sentence-picture and semantic memory verification. *Cognitive Psychology, 6*, 216–236.

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

Kanazawa, M. (2013). Monadic quantifiers recognized by deterministic pushdown automata. In M. Aloni, M. Franke, & F. Roelofsen (Eds.), *Proceedings of the 19th Amsterdam colloquium* (pp. 139–146).

Keenan, E., & Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy, 9*, 253–326.

Keenan, E. L., & Paperno, D. (2017). Overview. In D. Paperno, & E. L. Keenan (Eds.), *Handbook of quantifiers in natural language: Volume ii* (pp. 995–1004). Cham: Springer.

Kleene, S. C. (1951). *Representation of events in nerve nets and finite automata. Technical Report RM-704*. U.S. Air Force /RAND Corporation.

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

Knowlton, T., Hunter, T., Odic, D., Wellwood, A., Halberda, J., Pietroski, P., & Lidz, J. (2021). Linguistic meanings as cognitive instructions. *Annals of the New York Academy of Sciences, 1500*(1), 134–144.

Kounios, J., & Holcomb, P. (1992). Structure and process in semantic memory: Evidence from event-related brain potentials and reaction times. *Journal of Experimental Psychology: General, 121*(4), 459–479.

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

Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research, 44*(1), 27–46.

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of most. *Natural Language Semantics, 19*, 227–256.

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

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

Matthewson, L. (2001). Quantification and the nature of crosslinguistic variation. *Natural Language Semantics, 9*, 145–189.

McMillan, C. T., Clark, R., Moore, P., Devita, C., & Grossman, M. (2005). Neural basis for generalized quantifier comprehension. *Neuropsychologia, 43*(12), 1729–1737.

McMillan, C. T., Clark, R., Moore, P., & Grossman, M. (2006). Quantifier comprehension in corticobasal degeneration. *Brain and Cognition, 62*(3), 250–260.

McMillan, C. T., Coleman, D., Clark, R., Liang, T.-W., Gross, R. G., & Grossman, M. (2013). Converging evidence for the processing costs associated with ambiguous quantifier comprehension. *Frontiers in Psychology, 4*, 153.

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

Mock, J., Huber, S., Bloechle, J., Dietrich, J. F., Bahnmueller, J., Rennig, J., Klein, E., & Moeller, K. (2018). Magnitude processing of symbolic and non-symbolic proportions: An fMRI study. *Behavioral and Brain Functions, 14*(1).

Morgan, B., Gross, R. G., Clark, R., Dreyfuss, M., Boller, A., Camp, E., Liang, T. W., Avants, B., McMillan, C. T., & Grossman, M. (2011). Some is not enough: Quantifier comprehension in corticobasal syndrome and behavioral variant frontotemporal dementia. *Neuropsychologia, 49*(13), 3532–3541.

Moschovakis, Y. N. (2006). A logical calculus of meaning and synonymy. *Linguistics and Philosophy, 29*, 27–89.

Mostowski, M. (1998). Computational semantics for monadic quantifiers. *Journal of Applied Non-Classical Logics, 8*, 107–121.

Muskens, R. (2005). Sense and the computation of reference. *Linguistics and Philosophy, 28*, 473–504.

Nieuwland, M. S. (2016). Quantification, prediction, and the online impact of sentence truth-value: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(2), 316–334.

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language, 85*(2), 203–210.

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

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG,EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011*.

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

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of 'most': Semantics numerosity and psychology. *Mind and Language, 24*, 554–585.

Pietroski, P., Lidz, J., Hunter, T., Odic, D., & Halberda, J. (2011). Seeing what you mean, mostly. In J. Runner (Ed.), *Experiments at the interfaces, volume 37 of syntax and semantics* (pp. 181–217). Leiden: Brill.

Politzer-Ahles, S., Fiorentino, R., Jiang, X., & Zhou, X. (2013). Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification. *Brain Research, 1490*, 134–152.

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

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *TRENDS in Cognitive Sciences, 11*, 251–257.

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

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

Shetreet, E., Chierchia, G., & Gaab, N. (2014). When three is not some: On the pragmatics of numerals. *Journal of Cognitive Neuroscience, 26*(4), 854–863.

Spychalska, M., Kontinen, J., Noveck, I., Reimer, L., & Werning, M. (2019). When numbers are not exact: Ambiguity and prediction in the processing of sentences with bare numerals. *Journal of Experimental Psychology: Learning Memory and Cognition, 45*(7), 1177–1204.

Spychalska, M., Kontinen, J., & Werning, M. (2016). Investigating scalar implicatures in a truth-value judgement task: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience, 31*, 817–840.

Steinert-Threlkeld, S., & Szymanik, J. (2020). Learnability and semantic universals. *Semantics and Pragmatics, 12*(4).

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

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

Szymanik, J., & Thorne, C. (2017). Exploring the relation between semantic complexity and quantifier distribution in large corpora. *Language Sciences, 60*, 80–93.

Szymanik, J., & Zajenkowski, M. (2010a). Comprehension of simple quantifiers: Empirical evaluation of a computational model. *Cognitive Science, 34*(3), 521–532.

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

Szymanik, J., & Zajenkowski, M. (2011). Contribution of working memory in parity and proportional judgments. *Belgian Journal of Linguistics, 25*, 176–194.
Talmina, N., Kochari, A., & Szymanik, J. (2017). Quantifiers and verification strategies: Connecting the dots. In A. Cremers, T. van Gessel, & F. Roelofsen (Eds.), *Proceedings of the 21st Amsterdam colloquium* (pp. 465–473).
Tichý, P. (1969). Intension in terms of turing machines. *Studia Logica, 24*, 7–21.
Tomaszewicz, B. (2011). Verification strategies for two majority quantifiers in polishI. Reich, E. Horch, & D. Pauly (Eds.). . In *Proceedings of sinn und Bedeutung, 15* pp. 597–612).
Urbach, T. P., DeLong, K. A., & Kutas, M. (2015). Quantifiers are incrementally interpreted in context, more than less. *Journal of Memory and Language, 83*, 79–96.
Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language, 63* (2), 158–179.
van Benthem, J. (1986). *Essays in logical semantics*. Netherlands: Springer.
van de Pol, I., Steinert-Threlkeld, S., & Szymanik, J. (2019). Complexity and learnability in the explanation of semantic universals of quantifiers. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual conference of the cognitive science society* (pp. 3015–3021). Montreal, QB: Cognitive Science Society.
van Lambalgen, M., & Hamm, F. (2005). *The proper treatment of events*. Malden: Blackwell.
Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning Memory, and Cognition, 25*, 394–417.
Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*, 176–190.
van Rooij, I., & Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychological Inquiry, 31*(4), 321–325.
van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science, 16*(4), 682–697.
van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and intractability: A guide to classical and parameterized complexity analysis.* Cambridge: Cambridge Univeristy Press.
Vissers, C. T. W. M., Kolk, H. K. J., van de Meerendonk, N., & Chwilla, D. J. (2008). Monitoring in language perception: Evidence from ERPs in a picture-sentence matching task. *Neuropsychologia, 46*, 967–982.
Wassenaar, M., & Hagoort, P. (2007). Thematic role assignment in patients with Broca's aphasia: Sentence-picture matching electrified. *Neuropsychologia, 45*, 716–740.
Yang, H., Laforge, G., Stojanoski, B., Nichols, E. S., McRae, K., & K”ohler, S. (2019). Late positive complex in event-related potentials tracks memory signals when they are decision relevant. *Scientific Reports, 9*, 9469.
Zajenkowski, M., & Szymanik, J. (2013). MOST intelligent people are accurate and SOME fast people are intelligent Intelligence, working memory, and semantic processing of quantifiers from a computational perspective. *Intelligence, 41*(5), 456–466.
Zajenkowski, M., Szymanik, J., & Garraffa, M. (2014). Working memory mechanism in proportional quantifier verification. *Journal of Psycholinguistic Research, 43*(6), 839–853.
Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162–185.