In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), Proceedings of the 37th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Exploring the processing costs of the *exactly* and *at least* readings of bare numerals with event-related brain potentials

Maria Spychalska* (maria.spychalska@rub.de), Jarmo Kontinen* (jarmo.kontinen@rub.de), Ira Noveck (noveck@isc.cnrs.fr), Ludmila Rösch* (ludmila.roesch@rub.de), Markus Werning* (markus.werning@rub.de)**
**Institute of Philosophy II, Ruhr University Bochum*
***L2C2 CNRS Universit Lyon 1 - Institut des Sciences Cognitives*

Abstract

Bare numerals (e.g. *two*) seem to be ambiguous between two readings: the *exactly* and the *at least* reading. We present an ERP study that explores this issue. We show that the pattern of the ERPs elicited by critical nouns in sentences with unembedded bare numerals depends on the participant's choice of the reading of the numeral. For those responders who consistently apply the *exactly* reading in their truth-value judgement, sentences that are true only under the *at least* reading are associated with a sustained negativity effect compared to sentences that are true also under the *exactly* reading. However, no such effect is evident for the responders who apply the *at least* interpretation. We argue that this result falsifies the *exactly-theory* of numerals and speaks in favor of the ambiguity account. **Keywords:** bare numerals; *at least* reading; *exactly* reading; truth-value judgment; N400 effect; sustained anterior negativity

Introduction

One might think that no linguistic expressions are more precise than numerical expressions. For instance, saying that

(1) *Two roses are red.*

should leave no doubts regarding the number of red roses. Yet, such sentences as (1), where a bare numeral (*two*) occurs as a part of the quantifier phrase, seem to be ambiguous between various readings. Consider having a bouquet of five red roses. Is then sentence (1) true or false? If *two* is interpreted as *exactly two* (*exactly* reading), then (1) will be considered false in the given scenario. However, *two* may be interpreted as *at least two* (*at least* reading), in which case (1) will be considered true if there are more than two red roses. Since the *at least* reading implies only the lower bound for numerals, whereas the *exactly* reading implies both the lower and the upper bound, the first one is often referred to as the *weak* or *one-sided* reading and the latter is referred to as the *strong* or *two-sided* reading.

The choice between these two readings seems to be also dependent on complex interactions between contextual factors and linguistic structure. For instance, the preferred interpretation of sentence (2a) is that Anna has exactly two daughters, not more. In contrast, the common understanding of (2b) is that at least two essays are a condition for passing the course.

(2) a. *Anna has two daughters.*
b. *To pass the course one needs to write two essays.*
c. *If you have two children, you do not qualify for a tax refund.*

Moreover, there are contexts in which numerals invite so-called *at most* readings. Thus, (2c) is usually interpreted that if you have no more than two children, you do not qualify for a tax refund. Yet, since it is questionable whether the *at most* readings are at all available in unembedded contexts, in our current study we only consider the *exactly* and *at least* readings.

Up to date there is no agreement regarding the right semantic theory of bare numerals. According to the classical, neogricean view the *at least* reading constitutes the basic and literal meaning for numerals, whereas the *exactly* reading results from pragmatic strengthening by means of scalar implicature (Horn, 1992; Levinson, 2000; Schulz & Van Rooij, 2006). This strengthening is then similar as in the case of the quantifier *some*, whose semantic existential meaning is strengthened to *some but not all*. In short, the neogricean approach assumes that numerals are scalar terms and can be ordered on a linguistic scale according to their semantic strength: ⟨*one*, *two*, *three*, etc.⟩. Thus, sentence (1) literally means that there are at least two red roses, however, if the speaker knew that there were more than two red roses (e.g. three), then she should have provided a more informative statement with an appropriate stronger numeral (*Maxim of Quantity*). Based on such reasoning, (1) is taken to imply that there are not more than two red roses.

This standard view has been questioned on various grounds. First, numerals allow the strong reading in syntactic environments in which scalar implicatures are generally considered unavailable (e.g. downward entailing contexts) (Horn, 1992; Breheny, 2008). Second, developmental data suggest that children acquire the *exactly* reading of numerals earlier that the standard scalar implicatures (Noveck, 2001; Papafragou & Musolino, 2003; Hurewitz et al., 2006). Third, whereas the cognitive load in computing the scalar implicature of *some* is larger than in computing the existential meaning (Bott & Noveck, 2004; Bott et al., 2012), the *exactly* reading of numerals seems to be less demanding compared to the *at least* reading (Marty et al., 2013, 2014).

In the light of the shortcomings of the traditional view, various other approaches have been proposed. According to Carston (1988, 1998), numerals are underspecified and can be interpreted under either the *exactly*, *at least* or *at most* reading, and the interpretation of a particular occurrence of a numeral is fully determined by context. Breheny (2008) argues that the literal meaning of numerals is constituted by the *exactly* reading, whereas both the *at least* and *at most* readings are results of pragmatic processes. Accordingly, example (2b) means that one needs to write *exactly two* essays to pass the course but we arrive at the *at least* interpretation based on additional contextual assumptions, e.g. that the usual requirements for passing a course presuppose only a minimum number of essays and do not put any constraints on the maximum number. In a similar manner the *at most* interpretation of (2c) is based on the interaction of the *exactly* meaning of *two* and our knowledge of how the tax system works (Breheny, 2008; Spector, 2013). Finally, according to this *exactly-theory*, the *at least* reading in the case of unembedded numerals such as (1) is a result of a *pragmatic weakening* mechanisms of the strong reading: The domain of the quantifier can be implicitly restricted, which means that (1) can be interpreted as *Exactly two roses that are P are red* (where *P* is some contextually assumed property), therefore, (1) turns out true if there are three red roses.

In contrast, Geurts (2006) proposes, that numerals, when used as quantifiers, are lexically ambiguous between two distinct lexical entries corresponding to the *at least* and *exactly* readings. In this approach the *exactly* reading is still more basic, whereas the *at least* lexical entry can be obtained by a number of type-shifting operations. A different version of the ambiguity account has been offered by Spector (2013). Building on the grammatical account of scalar implicatures by Chierchia et al. (2012), Spector argues that albeit numerals are not lexically ambiguous, they give rise to ambiguities by means of interaction between the lexical entry and other operators, such as the exhaustification operator (*exh*). Yet, in unembedded contexts numerals are also argued to have a strong preference for being in the scope of *exh*, thus inviting the *exactly* reading.

In our study we tried to shed more light on this debate by investigating the processing of sentences with bare numerals with event-related brain potentials. We tested which of the two readings is more silent in unembedded contexts, such as *Two pictures contain stars*, and how the choice of reading (measured by the intuitive truth-value judgments given by the participants) modulates the ERPs evoked by critical nouns downstream the quantifier phrase, when the choice of a particular reading of the numeral makes those nouns more or less expected. We used a sentence-picture verification paradigm and asked our subjects to evaluate sentences with bare numerals in scenarios in which they were true according to the *exactly* reading, true according to the *at least* reading (but false under the *exactly* reading), or false under both readings. Based on the existing literature we expected that the critical *at least* cases should receive mixed truth-value judgments (Marty et al., 2013, 2014), where their evaluation as false indicates the *exactly* interpretation, whereas the accepting response indicates the *at least* interpretation.

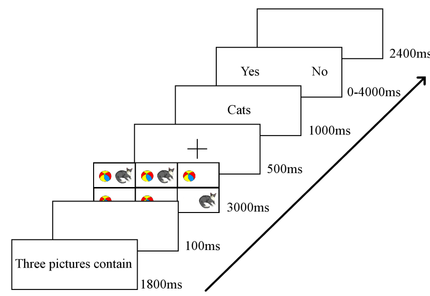
Experimental design

The experiment was conducted in German and had a form of a sentence-picture verification task. Participants were asked whether sentences of form (3) were true with respect to visual scenarios consisting of sets of six pictures and depicting two categories of objects (Xs and Ys) in different quantities.

(3) *N pictures contain X.*

As the numeral *N* we used either *three* (*drei*) or *four* (*vier*) in all experimental sentences. In each trial, subjects (i) were first presented with the phrase containing the numeral *Three/four pictures contain* (*Drei/Vier Bilder enthalten*); (ii) next they saw a scenario; and (iii) finally they saw the critical noun *X* (Figure (1)). They were asked to respond at the end of each trial (after the final word disappeared) whether the sentence was true with respect to the presented scenario. ERPs were measured on the onset of the critical noun *X*.

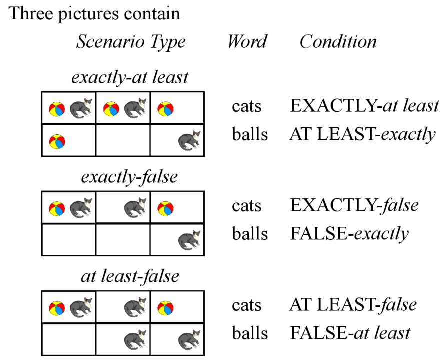
Figure 1: Time-course of experimental trials.



There were six experimental conditions, determined by the number of Xs presented in the scenario, i.e. the number of objects that were mentioned at the end of the trial, and the number of the non-mentioned objects (Ys). The number of Xs determined the *evaluation condition* of the sentence and was either smaller than the numeral *N* occurring in the sentence (*FALSE condition*), exactly *N* (*EXACTLY condition*), or larger than *N* (*AT LEAST condition*). However, the non-mentioned alternative was also represented in the *exactly*, *at least* or *false* way with respect to the numeral *N* and created a *context scenario* for the sentence evaluation. For instance, in condition *EXACTLY-false* the number of the mentioned objects X was exactly *N* and the number of the non-mentioned objects Y was smaller then *N*. In total there are 3×3 possible combinations, however, we eliminated those conditions in which both X and Y fell under the same evaluation. In the end there were (i) two *EXACTLY* conditions: *EXACTLY-at least* and *EXACTLY-false*; (ii) two *AT LEAST* conditions: *AT LEAST-exactly* and *AT LEAST-false*; and (iii) two *FALSE* conditions *FALSE-exactly* and *FALSE-at least*. However, there was no *EXACTLY-exactly*, *AT LEAST-at least* or *FALSE-false* condition. Figure (2) presents all six conditions as combinations of *scenario types* and critical words: Two conditions are always associated with one *scenario type* that is determined by the combination of the mode (*at least/exact/false*) in which both X and Y are represented in a picture with respect to given numeral.

Our planned comparisons primarily involved contrasting the respective evaluation conditions for the same scenario

Figure 2: Six conditions as combinations of the numeral, scenario types and critical words.



types: *EXACTLY-at least* with *AT LEAST-exactly*, *AT LEAST-false* with *FALSE-at least* and *EXACTLY-false* with *FALSE-exactly*. In this way we could reciprocally contrast the evaluation conditions while the context proceeding the critical noun onset, i.e. the initial sentence phrase plus the scenario type, was identical.

A conservative interpretation of the *exactly-theory* predicts that the ERPs for the *AT LEAST-exactly* cases should be more negative in the N400 time-window compared to the *EXACTLY-at least* cases across the tested population. Admittedly, one could argue that some participants might apply the *at least* reading in their truth-value judgment due to the possibility of the *pragmatic weakening* mechanism. However, even for those participants the *AT LEAST-exactly* condition should be associated with a larger N400 than the *EXACTLY-at least* condition, or at least we should observe some other ERP effect (e.g. P600) indicating the relevant pragmatic processes taking place. In contrast, if the *at least* reading is the literal interpretation, as predicted by the neogricean approach, the N400 ERPs for critical nouns in the *AT LEAST-exactly* and *EXACTLY-at least* conditions should not differ, since from a semantic point of view in both cases the evaluated sentences are equally true. Yet, it is a known result that people tend to base their truth-value judgments not only on the literal semantic meaning but often integrate the implicature into the sentence's truth-conditional content. Spychalska et al. (2014) show that for such *pragmatic responders* N400 ERPs associated with underinformative *some*-sentences are larger than those associated with informative sentences. However, this result can be in fact taken as evidence against the traditional gricean view on scalar implicatures and in favor of those theories in which scalar implicatures can be computed in par with the compositional content of the sentence (e.g. Chierchia et al., 2012). A similar result in the case of numerals, i.e. a correlation between the behavioral truth-value judgments and the ERP patterns, would not support neogricean approach, but rather the lexical ambiguity view or the grammatical account in which ambiguity is a result of the exhaustification operator.

The predictions of the ambiguity view are the following: If numerals are ambiguous between the two readings, then the participants' choice of the reading of the numeral (reflected in the truth-value judgments for the critical *AT LEAST* cases) should correlate with the ERP patterns. This means that the choice of the *exactly* reading should trigger a larger negativity for the *AT LEAST-exactly* compared to the *EXACTLY-at least* condition, whereas the choice of the *at least* reading should result in a larger negativity for the *FALSE-at least* compared to the *AT LEAST-false* cases.

Since in both versions of the ambiguity view the *exactly* reading is still considered to be the preferred reading in unembedded contexts, one could argue that this preference should leave a mark in the ERP pattern. Therefore, it makes it perhaps difficult to contrast the ambiguity view and the *exactly-theory*. However, one should note that under the bilateral *exactly-theory* the *at least* reading is not a genuine reading. Thus, one could argue that the *exactly-theory* has the following prediction: For those subjects for whom no pragmatic weakening is evident (who respond according to the *exactly* reading), we expect that the processing of the *AT LEAST-false* and *FALSE-at least* cases should be alike. However, if the *at least* reading is a genuine, just less preferred reading—as predicted by the ambiguity view, we should be able to observe a mark in the ERPs indicating that the *AT LEAST* cases constitute a weaker violation than the *FALSE* cases also for the "exact" responders.

Materials

We used 240 unique German nouns to construct two different lists of 120 ordered pairs $\langle n_1, n_2 \rangle$. All words were used in their plural form, were two-syllabic and had a length of 4 to 9 characters; compound nouns were excluded. The word frequency value was checked in the *Wortschatz Leipzig* corpus (http://wortschatz.uni-leipzig.de/)[1], and was kept between 8 and 17 (moderate frequent words). All nouns denoted concrete objects, that are easy to identify in a picture and are well-known to an average German speaker. The two lists were fully exclusive with respect to the word-pairs. Thus, each word occurred exactly once in each list (so twice in total), but was combined in each of the lists with a different word. Within each pair the words were matched with respect to their length (maximal character difference was 4) and frequency (maximal value difference was 4), as well as for their semantic similarity value (LSA) (Landauer et al., 1998). We estimated the LSA value based on the English translations of the singular German nouns by using the http://lsa.colorado.edu/ server. The words were matched so that the LSA values for all pairs did not exceed 0.3. In order to eliminate heavy outliers we computed the standardized values (z-scores) and recombined the words in such a way that at the end the z-scores for both lists were between -2.5 and 2.5.

A picture of a corresponding object was created for each noun and for each pair of nouns using free clipart images as well as Adobe Photosphop. For presenting the stimuli we used NBS Presentation® Software. For each participant the program generated a unique stimuli list in a pseudo-random

[1]The frequency *value* v of a word w is equal to log_2 of the quotient of the frequency of the word "der" and the frequency of the word w in corpus

manner from the predefined conditions list, the two lists of noun pairs and pictures database. There were 240 experimental trails (40 per condition) and 96 filler trials. The filler trails used quantifiers *all* (*alle*), *no* (*keine*), *two* (*zwei*) or *five* (*fünf*) and were created based on a set of 48 nouns, which were less strictly controlled for frequency or length than the nouns used in the experimental trails.

Participants

Forty-five (twenty-five woman) members of the Ruhr-University Bochum were recruited for the experiment (age: 18-42, mean: 24.8 SD: 5.4). They were reimbursed for their participation. All participants were monolingual German native speakers, had at least a secondary degree (German *Abitur*), normal or corrected to normal vision, no history of psychological or neurological problems, and were right-handed. Two people were excluded from the EEG analysis due to excessive noise in their EEG data.

Procedure

Upon arrival all our participants signed a written consent of participation including a statement concerning their vision, medication, neurological or psychiatric history. They filled in the *Edinburgh Handedness Inventory* test and were screened using the WAIS test for logic fluid intelligence, the digit span working memory test and the *AQ Questionnaire*. Additionally they were also tested using a modified version of the Reading Span Memory Test, German version (Noort et al., 2008). The measurement was conducted in a dim, electrically and acoustically isolated cabin. Subjects were seated in front of a computer screen and a response pad with two designated buttons. The experiment started with a short instruction followed by an exercise session consisting of five example trials. No feedback was given throughout the experiment and subjects were asked to follow their intuition in the truth-value judgment task. The time-course of experimental trials is presented in Figure (1).

EEG recording and data processing

EEG was recorded from 64 active electrodes held on the scalp by an elastic cap, with a BrainAmp acticap EEG recording system. AFz served as the ground electrode and FCz—as the physical reference. Four electrodes (FT9, FT10, P09, PO10) were reprogrammed and used for controlling both vertical (above and below the right eye) and horizontal (on the right and left temple) eye-movements (EOG electrodes). The EEG was recorded with a sampling rate of 500 Hz and a low cut-off filter of 0.01. Impedance was kept below $5k\Omega$.

The EEG data were processed using Brain Vision Analyzer 2.0 software. We applied an off-line high cut-off filter at 30 Hz, 12 dB/oct. Automatic raw data inspection rejected all trials with the absolute amplitude difference over $200\mu V/200ms$, or with the activity lower than $0.5\mu V$ in intervals of at least $100ms$. The maximum voltage step allowed was $50\mu V/ms$. Eye blinks were corrected using an independent component analysis. The data was off-line re-referenced to the average of linked mastoids comprising of TP9 and TP10. Segments from $200ms$ pre-target onset until $1000ms$ post-onset were separately extracted and averaged for every subject and every condition. Baseline correction used the $200ms$ interval preceding the onset of the stimulus. All segments with any remaining physical artifacts (including the amplitude lower than $-90\mu V$ or higher than $90\mu V$) were excluded before averaging. The minimum number of segments that was preserved in each condition was 23 out of 40.

Results

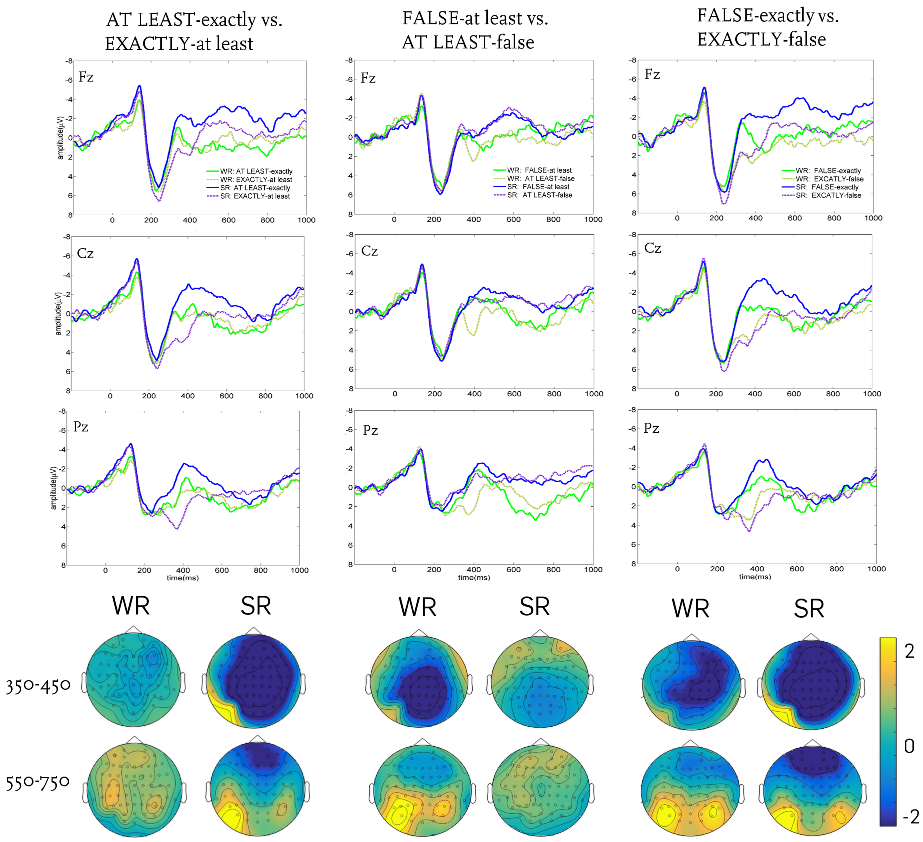
Behavioral results

The analysis of truth-value judgements indicated that, for both critical *AT LEAST* cases, i.e. *AT LEAST-exactly* and *AT LEAST-false*, our participants were consistent in their choice of either the *exactly* or the *at least* reading. Based on the mean of the proportion of the *exactly/at least* readings in both *AT LEAST* conditions, we divided our participants into two groups of responders. Most participants ($N = 30$, 66.7%; two were later excluded from the analysis due to noise in their EEG) consistently said "false" to the *AT LEAST* cases, i.e. applied the *exactly* reading of the numeral, and are henceforth referred to as *strong readers* (the proportion of *exactly* readings varied between 88.46% and 100% in all *AT LEAST* trails). The remaining responders ($N = 15$, 33.3%), who consistently said "true" to the *AT LEAST* cases, are called *weak readers* (between 88.75% and 100% of the *at least* readings). Accordingly, we defined accuracy for the *AT LEAST* cases based on the response profile: the rejecting response was correct for the *strong readers* and the accepting response was correct for the *weak readers*. The accuracy in all remaining conditions was at the ceiling level. There were no significant differences between the two groups in any of the measured cognitive or personality tests.

EEG results

The visual inspection of grand averages allowed us to conclude that *strong* and *weak readers* obtained a different pattern of the ERP waveform. For the statistical analysis of the EEG data we used the Matlab Fieldtrip package. We performed a non-parametric statistical procedure called cluster-based permutation test (Maris & Oostenveld, 2007). For each subject the ERPs were averaged across trials in the compared conditions, in the epoch of $0-1000$ ms post-onset and for all channels. The data-points (time $\times$ channel) between the sets were compared by a two-tailed dependent t-test. The significantly different ($\alpha = 0.025$) data-points were then clustered according to the time-spatial adjacency. The cluster-level statistics were calculated by taking the sum over the t-values for each cluster. The cluster-level p-values were evaluated with a Monte Carlo simulation: For each subject the ERP averages were randomly swapped between the two conditions. The cluster-level statistics were computed again and the maximum of the cluster-level statistics was taken as the

Figure 3: The comparison of grand averages of all conditions, for *weak readers* (WR) and *strong readers* (SR). Topographical maps of the effects (difference curves of the compared conditions) in all three comparisons in the time-windows of 350-450 ms and 550-750 ms. Each column presents one comparison.



test statistics. This procedure was repeated 10000 times and the p-values of the observed cluster-level statistics were estimated as the proportion of permutations that resulted in a higher test-statistics than the observed one.

For both groups the control comparison between falsity and unambiguous truth, i.e. *FALSE-exactly* and *EXACTLY-false* conditions, resulted in significant negativity effects. For the *strong readers* the effect had a form of a sustained negativity lasting from around 282 till 788 ms post-onset[2], which had first global and later frontal topographical distribution. For the *weak readers* the negativity effect was only significant around the N400 time-window (286 – 438 ms), had a global distribution and was followed by a marginally significant posteriorly distributed P600 (524 – 712 ms). (See Table 1 for the time-windows and p-values of all the observed significant clusters).

The main comparison between the experimental conditions *AT LEAST-exactly* and *EXACTLY-at least* resulted in a sustained negativity effect for the *strong readers* (162 – 610 ms)—the effect was global around the N400 time-window and extended as a sustained anterior negativity. For the *weak readers* there were no significant effects for this contrast. A contrastive result was obtained for the comparison between *FALSE-at least* and *AT LEAST-false* conditions: Whereas the *weak readers* showed a significant (centro-posterior) N400 effect (362 – 446 ms), for the *strong readers* the observed N400 cluster (400 – 492 ms) was only marginally significant ($p < .066$).

To statistically explore the interaction between the applied reading and the observed ERP effects, we calculated the difference curves for both groups in each of the comparisons. These difference were averaged for each channel in time-bins of 100 ms starting from 200 ms post-stimulus onset until 800 ms. We compared the averages in each 100 ms time-bin between the two groups with a cluster-based permutation test that used an independent t-test and 10000 permutations for the estimation of a p-value. For each 100 ms the analysis produced as an output a cluster of channels where one group had a more negative/positive effect in the given comparison than the other one. The analysis revealed that the difference between the *AT LEAST-exactly* and *EXACTLY-at least* conditions was more negative for the *strong readers* in three consecutive time-windows: 200 – 300 ms ($p = .024$, clustering mainly on frontal and central channels), 300 – 400 ms ($p < .0001$, globally), 400 – 500 ms ($p = .009$, most pronounced on fronto-central channels). However, there was no significant between-group difference in the effect observed in the comparison *FALSE-at least* and *AT LEAST-false* in any of the time-windows. It is also interesting that the difference between conditions *FALSE-exactly* and *EXACTLY-false* was

[2] The time window indicates the maximal latency of the respective cluster.

Table 1: The positive and negative clusters for all comparisons, for *weak* and *strong readers*. Marginally significant clusters are also reported. The empty spaces indicate the lack of significant clusters. Additionally, the results of the permutation tests performed on the whole group of subjects. (Abbreviations: *at=at least*, *e=exactly*, *f=false*.)

			AL-e vs. E-al	*F-e vs. E-f*	*F-al vs. AL-f*
Strong Readers	Neg. clusters	time (ms)	162 – 610	282 – 788	400 – 492
		sig.	$p < .0002$	$p < .0001$	$p < .066$
		region	*global*	*global*	*centro-posterior*
			sust. anterior	*sust. anterior*	
Weak Readers	Neg. clusters	time (ms)		286 – 438	362 – 446
		sig.		$p < .016$	$p < .049$
		region		*global*	*centro-posterior*
	Pos. clusters	time(ms)		524 – 712	
		sig.		$p < .051$	
		region		*posterior*	
Both	Neg. clusters	time (ms)	188 – 570.	230 – 808	334 – 516
		sig.	$< .0005$	$p < .0001$	$p < .0125$
		region	*global*	*global*	*centro-posterior*
				sust. frontal	
	Pos. clusters	time(ms)		516 – 790	
		sig.		$p < .0442$	
		region		*posterior*	

more negative for the *strong readers* in the time-window of 400 – 500 ms on centro-parietal channels ($p = .004$), and on the parietal sites in the time window of 500 – 600 ms ($p = .013$).

Discussion

The results of our experiment support the ambiguity account of bare numerals and speak against the *exactly-theory*. We have shown that the pattern of the ERPs elicited by critical nouns in sentences with unembedded bare numerals depends on the participant's choice of the reading of the numeral. In the identical scenario type critical nouns in the *AT LEAST* condition elicited a large N400 effect extending into a sustained anterior negativity when compared to critical nouns in the *EXACTLY* condition only for those subjects who applied the *exactly* reading in their truth-value judgement. In contrast, those responders who applied the *at least* reading in their truth-value judgment, there were no ERP differences in this comparison. This result speaks against the bilateral *exactly* semantics of numerals. If the literal meaning was constituted by the *exactly* reading, we would expect some ERP effect for the *AT LEAST-exactly* vs. *EXACTLY-at least* comparison also for those participants who responded according to the *at least* reading, signalling the necessary *pragmatic weakening* mechanisms.

The result for the *FALSE-at least* vs. *AT LEAST-false* comparison can be also taken in favor of the ambiguity view and against the bilateral *exactly* semantics. The *exactly-theory* does not predict why the *strong readers* should show any N400 effect in this case—it does not explain why the *strong readers*' expectations for critical nouns in the *AT LEAST* cases should differ with respect to the unambiguously *FALSE* cases. However, we have observed that even for the *strong readers* there was a marginally significant N400 effect. Moreover, there were no significant differences between the *strong* and *weak readers* in the size of the effect in this comparison in any of the consecutive 100 ms time-windows. Thus, independently of the applied reading the *AT LEAST* cases elicited smaller N400 ERPs than the *FALSE* cases.

Our experiment provides evidence that in spite of the predominant character of the *exactly* reading in unembedded cases, the *at least* reading cannot be considered only a result of a pragmatic process. This is an important empirical result for the debate regarding the semantics of numerals.

Acknowledgments

This study was supported by the *Stiftung Mercator* Foundation.

References

Bott, L., Bailey, M., Todd, & Grodner, D. (2012, January). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, *66*(1), 123-142.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar implicatures. *Journal of Memory and Language*, *51*(3), 437-457.

Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantifed noun phrases. *Journal of Semantics*, *25*, 93-140.

Carston, R. (1988). Implicature, explicature, and truth-theoretic semantics. In R. Kempson (Ed.), *Mental representations: The interface between language and reality,* (p. 155-181). Cambridge University Press, Cambridge.

Carston, R. (1998). Informativeness, relevance and scalar implicature. In R. Carston & S. Uchida (Eds.), *Relevance theory: Applications and implications* (p. 179-236). Benjamins, John.

Chierchia, G., Fox, D., & Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.), *An international handbook of natural language meaning.* Mouton de Gruyter.

Geurts, B. (2006). Take "five": the meaning and use of a number word. In S. Vogeleer & L. Tasmowski (Eds.), *Non-definiteness and plurality* (p. 311-329). Benjamins, Amsterdam/Philadelphia.

Horn, L. R. (1992). Proceedings from the second conference on semantics and linguistic theory. In C. Barker & D. Dowty (Eds.), *Working papers in linguistics salt ii* (p. 163-192). The Ohio State University.

Hurewitz, F., Papafragou, A., Gleitman, L., & Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *Language learning and development*, *2(2)*, 77-96.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259-284.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Massachusetts Institute of Technology.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177-190.

Marty, P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua*, *133*, 152-163.

Marty, P., Chemla, E., & Spector, B. (2014). Phantom readings: the case of modified numerals. *Language, Cognition and Neuroscience*, *ahead-of-print*, 1-16.

Noort, M. Van den, Bosch, P., Haverkort, M., & Hugdahl, K. (2008). A standard computerized version of the reading span test in different languages. *European Journal of Psychological Assessment*, *24(1)*, 35-42.

Noveck, I. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78 (2)*, 165-188.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semanticspragmatics interface. *Cognition*, *86 (3)*, 253-282.

Schulz, K., & Van Rooij, R. (2006). Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy*, *29*, 205-250.

Spector, B. (2013). Bare numerals and scalar implicatures. In *Language and linguistics compass.* Wiley Online Library.

Spychalska, M., Kontinen, J., & Werning, M. (2014). Electrophysiology of pragmatic processing: Exploring the processing cost of the scalar implicature in the truth-value judgment task. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society. austin, tx: Cognitive science society.*