Reading times show effects of contextual complexity and uncertainty in comprehension of German universal quantifiers

Fabian Schlotterbeck & Petra Augurzky*

Abstract. We report three experiments, in which we combined self-paced reading with picture-sentence verification to test how reading times are affected by meaning-related processes. In particular, we investigated German sentences containing the universal quantifier *alle* ("all") and examined how restrictive processes incrementally interact with other aspects of quantifier meaning, comparably to previous studies using other methods. Our results show that reading times were sensitive towards a match between context and sentence meaning and also towards an interaction between picture complexity and task demands. The results also point to the need for integrated processing models that combine refined notions of the relation between memory and expectations, on the one hand, with assumptions about adaptive processes and about representations involved in compositional interpretation, on the other.

Keywords. language comprehension; compositional-semantic processing; quantifier restriction; self-paced reading; expectation-based processing

1. Introduction. Current semantic and pragmatic theory offers detailed models of meaning-related processes for a wide range of linguistic phenomena. These models go beyond classical approaches in the sense that they not only intend to explain the compositional derivation of sentence meaning in general, but also focus on phenomena like incremental meaning composition (e.g. Bott & Sternefeld 2017), the complexity of meaning representations (e.g. Pietroski et al. 2009, Szymanik 2016) and contextual effects on the behavior of speakers and listeners (e.g. Frank & Goodman 2012, van Tiel et al. 2021). Despite these recent advances, relating predictions derived from semantic and pragmatic theory to processes during online comprehension remains an elusive goal in spite of the fact that theory-driven syntactic considerations have been implemented into models of on-line sentence comprehension for decades. This is especially surprising as highly comparable linking hypotheses could be developed on the basis of recent semantic and pragmatic models. For example, one could assume that complex meaning representations are generally avoided, or that highly expected sentence continuations lead to facilitation during incremental processing. We attempt to bridge this gap by studying how complexity and uncertainty in sentence meaning affect on-line sentence comprehension.

In the current experiments, we combined self-paced reading with picture-sentence verification to test how reading times are affected by meaning-related processes. In particular, we examined how restrictive processes incrementally interact with other aspects of quantifier meaning, comparably to previous studies using other methods (Augurzky et al. 2017, 2019, Bott et al. 2019).

*We would like to thank Roman Dick, Hening Wang & Margarethe van Liempt for help with programming and data preparation. Authors: Fabian Schlotterbeck, University of Tübingen (fabian.schlotterbeck@uni-tuebingen.de) & Petra Augurzky, Goethe University Frankfurt (augurzky@lingua.uni-frankfurt.de). FS received funding from the Baden-Württemberg Ministry of Science (MWK-BW) and the Federal Ministry of Education and Research (BMBF) as part of the Excellence Strategy of the German Federal and State Governments. PA received funding by Project B1 of the DFG-funded SFB 833, and the DFG-/AHRC-funded project IDEAlISM.

As an illustration of the basic phenomenon, consider the German sentences in (1), containing the universal quantifier *alle* ("all"), in the context of one of the pictures in Fig. 1.

(1) Alle Dreiecke sind blau — a. ..., die innerhalb .../ — b. ..., die außerhalb des Kreises sind.
All triangles are blue — ..., that inside .../ — ..., that outside of_the circle are

Each picture showed colored objects, e.g. triangles, placed inside or outside a container shape, e.g. a circle. In that kind of setting, a truth-value judgment is, in principle, possible on the color adjective *blue*, but it may have to be revised later on if further restriction is provided, e.g. in the form of a restrictive relative clause (cf. example (1-a) in the context of Fig. 1c where all and only the triangles inside the container shape are blue). Previous studies following Augurzky et al. (2017) repeatedly found evidence for an incremental truth evaluation on the adjective, but this effect was sensitive to the risk of a potential revision in the truth-value. In risky contexts, such as in Fig 1c, in which a change in truth-value was possible, evidence for non-incremental truth evaluation was found. This finding was initially interpreted as an indication of REVISION-SENSITIVITY: The processor plans ahead and only commits to an interpretation if there is no substantial risk to revise that interpretation later on. However, in subsequent studies, alternative accounts were also discussed, and we aim to compare these alternatives based on data from self-paced reading in the current study. In particular, we compare REVISION SENSITIVITY to two competing accounts. The first alternative is based on PRAGMATIC EXPECTATIONS. It assumes that comprehenders expect utterances that match the context, i.e. utterances that are true in the context (see e.g. Augurzky et al. 2019, Bott et al. 2019, Schlotterbeck et al. 2022; for discussion). Under this account, facilitation during processing is expected for words that allow for a high proportion of true vs. false continuations. The second alternative, which is also expectancy-based, explains modulation of processing difficulty in terms of PRIMING. Here the idea is that words denoting properties or concepts that are salient in the context lead to facilitation. Note that all of these accounts are inherently interrelated. For example, REVISION SENSITIVITY to some degree presupposes expectation-based processing (cf. Schlotterbeck et al. 2022; for discussion) and PRAGMATIC EXPECTATIONS may in part be driven by the salience of contextual properties (cf. Frank & Goodman 2012). Despite these relations, we aimed to distinguish as clearly as possible between these three accounts in the following experiments. Thus, we test a version of REVISION SENSITIVITY that is not based on specific expectations but rather on the mere possibility of revisions of truth values. Moreover, we take expectations that are based on salience rather than propositional meaning as evidence for PRIMING rather than for PRAGMATIC EXPECTATIONS.

2. Experiment 1. The purpose of the first experiment was to replicate an ERP experiment by Augurzky et al. (2017) using self-paced reading. Based on previous results, we expected to find facilitation for unambiguously true vs. false contexts but no such effects in risky contexts. The obtained results should serve as a baseline for further comparisons that are relevant to our three hypotheses, i.e. REVISION SENSITIVITY, PRAGMATIC EXPECTATIONS and PRIMING.

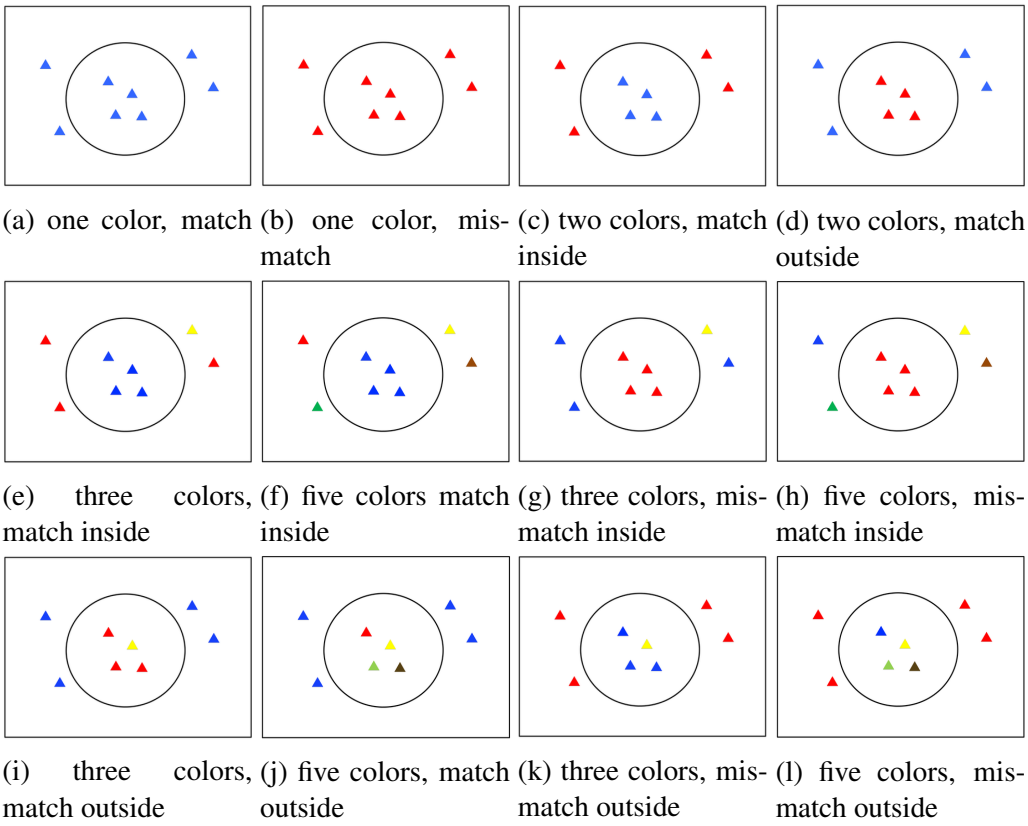
2.1. METHODS. In each trial of the current self-paced reading experiment, participants first inspected a picture context showing geometrical objects inside and outside of a container shape (e.g. one of the Figs. in 1a-d) and then read a universally quantified German sentence as in the examples in (1). Half of the sentences contained a restrictive relative clause, as in (1-a/b), which could lead

to a possible meaning change by introducing a subset reading (i.e. a reading where the quantificational domain is restricted to a subset of the geometrical objects, e.g. those outside the container shape). In total there were thus eight conditions in a 4 (CONTEXT) × 2 (PREPOSITION) design. For sentences following the simple, unicolored contexts (Fig. 1a/b), a truth-value judgment is possible already on the adjective (*blau*, 'blue'). By contrast, for bicolored contexts (Fig 1c/d), the judgment has to be delayed until the preposition (*innerhalb*, 'inside-of', *außerhalb*, 'outside-of') is encountered. Eight items per condition were constructed by combining various color adjectives (*blue*, *red*; *green*, *yellow*; *orange*, *purple*; *gray*, *black*), geometric shapes (triangles, pentagons, semicircles, hearts) and container shapes (rectangles, circles), yielding 64 experimental sentences. These were matched with 64 short filler sentences (as in (1), each presented twice). In addition, there were 16 catch trials, which prompted participants to make truth-value judgments about simple logical statements presented without contexts. In total this resulted in 144 trials that were distributed over four blocks using a Latin square design. Sentences were presented word-by-word (with punctuation displayed separately) using the moving-window technique, and participants performed a truth-value judgment task after each sentence. Fifty-nine German native speakers were recruited over the platform `prolific.co` and were paid for their participation. We excluded five participants from further analysis because they met one of the following exclusion criteria, which we used in all three experiments: word reading times were over 10 s in at least one trial; overall accuracy was below 70%; performance on catch trials was below 70%; or duration of the entire experiment was more than 3 standard deviations above the mean.

2.1.1. STATISTICAL ANALYSIS. For inferential statistics we used mixed-effects models (Bates et al. 2015) that were similar in the current and the following two experiments, with only a few differences due to adaptations to each experimental design. For accuracy, we used logit mixed-effects models, which included the fixed effects of CONTEXT, PREPOSITION and their interaction. RTs were trimmed individually for each participant and condition by removing extreme RTs that were shorter than 200 ms or longer than three standard deviations above the participant's mean in that condition. For RTs on the adjective, linear mixed-effects models were computed that predicted log-transformed RTs based on CONTEXT. For each of the experiments, custom contrasts for the multilevel factor CONTEXT were specified to test our hypotheses (see Table 1). They are explained in the results sections of the individual experiments. For RTs on the preposition, the factor PREPOSITION was also included, but for simplicity we focus on effects of truth values (i.e. specific interactions between CONTEXT and PREPOSITION) as they were predicted by REVISION SENSITIVITY. Other effects are only mentioned if theoretically relevant. For some effects we had derived directed predictions from our hypotheses. We thus calculated one-tailed t-tests for them and marked them as such in the results sections. In all models, by-participant and by-item random intercepts and slopes were included if they allowed for convergence and contributed to model fit. [1]

2.2. RESULTS. Across conditions, truth-value judgments were correct in the majority of cases ($94.0\% - 98.7\%$; see Fig. 2b). The highest accuracy was achieved in the false unicolored conditions, leading to a marginal effect of CONTEXT ($\chi^2(3) = 6.34, p = .096$), but, at the same time, no significant difference to any of the other context conditions was found ($|z| < 1.6$). Reading

[1] In the third experiment below, only random intercepts were included for participants due to the between design.



(a) one color, match (b) one color, mismatch (c) two colors, match inside (d) two colors, match outside

(e) three colors, match inside (f) five colors match inside (g) three colors, mismatch inside (h) five colors, mismatch inside

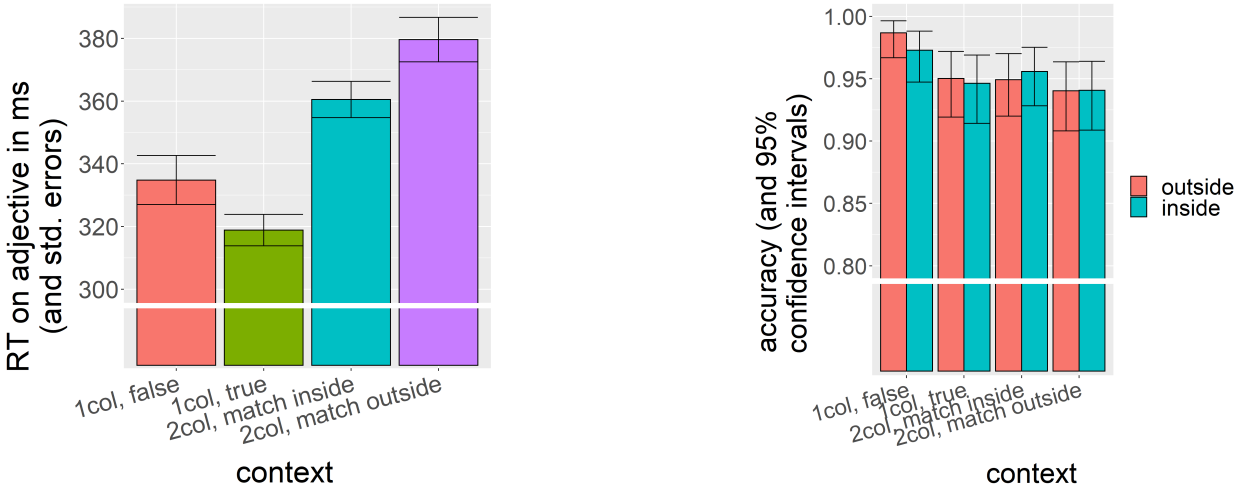
(i) three colors, match outside (j) five colors, match outside (k) three colors, mismatch outside (l) five colors, mismatch outside

Figure 1: Sample visual contexts (Exp. 1: a-d; Exp. 2: a-d, e & i ; Exp. 3: a-l)

Exp .1		Exp .2		Exp. 3					
#	Contrast	#	Contrast	#	Contrast	# contd.	Contrast	# contd.	Contrast
1	a, b vs c, d	1	a, b vs c-e, i	1	a, b vs c–l	6	e, f vs. g, h	9	i, j vs. k, l
2	a vs. b	2	a vs. b	2	a vs. b	7	g vs. h	10	k vs. l
3	c vs. d	3	c vs. d	3	c vs. d	8	e vs. f	11	i vs. j
		4	c, d vs. e, i	4	c, d vs. e–l				
		5	e vs. i	5	e–h vs. i–l				

Table 1: Contrasts for the factor CONTEXT used in the three experiments: Labels of subfigures in Fig. 1 are used as short references to context conditions

times at the adjective are shown in Fig. 2a. For the unicolored contexts, true conditions were read faster than false ones (one-tailed: $t = 1.66, p = .049$). For the bicolored contexts, a marginal difference between context where the matching color was inside vs. outside the container shape was found ($t = 1.7, p = .086$). Moreover, bicolored contexts led to significantly longer RT than unicolored ones ($t = 2.4, p = .017$). The latter effect was sustained over several regions and turned out to be reliable on the preposition as well ($\beta = 0.11, t = 5.81, p < .001$). Moreover, there was a truth-value effect for the bicolored contexts on the preposition (CONTEXT×PREPOSITION:



(a) Mean RT on adjective in Exp. 1

(b) Accuracy in Exp. 1

Figure 2: Results from Exp. 1

$\chi^2(1) = 4.07, p = .044$) which was due to shorter RT for the matching than mismatching bicolored contexts (significant for the preposition *inside*: means (sds): 332 (184) vs. 374 (299), $t = 2.7, p = .007$; but not *outside*: means (sds): 355 (184) vs. 350 (299), $t = -0.15, p = .007$).

2.3. DISCUSSION. The current self-paced reading results indicate that, comparable to ERPs (Augurzky et al. 2017), RTs are generally sensitive to the truth value of an utterance: We observed effects of truth values as soon as they could be unambiguously decided, i.e. on the color adjective in unicolored contexts and on the preposition in bicolored contexts. These effects may reflect a local truth evaluation and are thus in accord with REVISION SENSITIVITY. However, they are also in line with PRAGMATIC EXPECTATIONS as they may alternatively signal a facilitation for expected sentence continuations that allow for true descriptions of the context. Finally, these results are also compatible with PRIMING, as they could reflect facilitation effects for reading words denoting properties that were salient in the visual contexts. However, purely lexical PRIMING cannot explain the truth-value effect at the preposition, as the truth value at this point depends on both the preposition and the adjective. Thus, the salience of the matching preposition in the visual context cannot be the sole source of facilitation. We may instead assume that combined properties are primed (e.g. blue-inside), implying a rudimentary from of composition, or that the adjective is a cue for memory retrieval of the matching preposition. Under both of these assumptions, the hypothesis of PRIMING would look a bit more similar to PRAGMATIC EXPECTATIONS. The following two experiments were aimed to distinguish between these competing explanations.

On top of these potentially truth-value related effects, we observed a substantial slowdown for the bi- vs. unicolored picture contexts across several words. We interpret these effects as being related to memory demands: Whereas only one color term had to be remembered in the unicolored contexts, two colors and their positions were relevant to the experimental task in the bicolored contexts. Below, in the general discussion, we reflect on theoretical explanations of this effect. For now it is simply important to acknowledge this effect while we attempt to dissect the truth-value

related effects in the following two experiments.

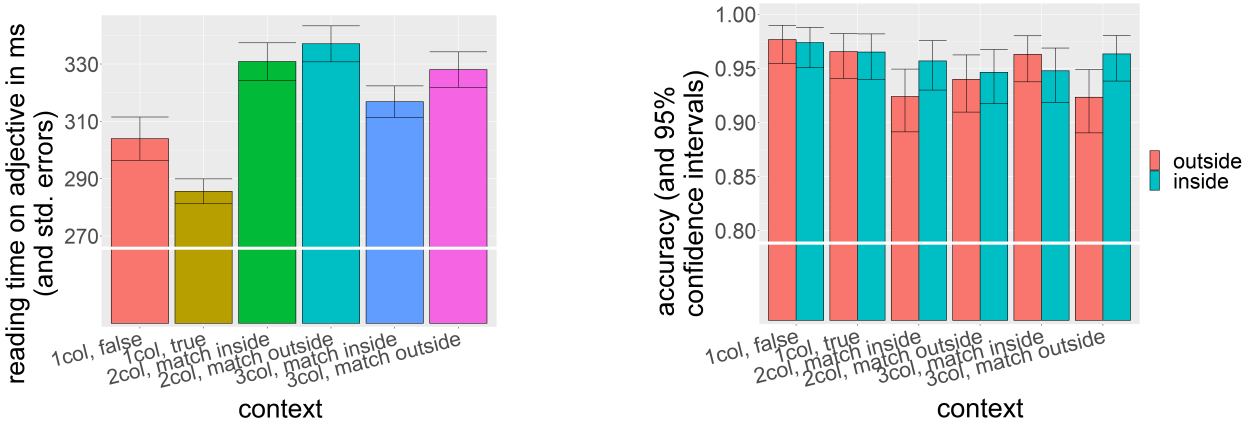
3. Experiment 2: Testing for expectation-based effects. In order to disentangle the two expectation-based accounts (i.e. PRIMING and PRAGMATIC EXPECTATIONS) from REVISION SENSITIVITY, we included another pair of visual contexts into the experimental design (see Fig. 1e/i). In these pictures, one element of the mismatching set of objects (inside or outside the container shape) had a different, third color that was not mentioned in the sentence. This change does not affect the predictions from REVISION SENSITIVITY: The newly added contexts still make the sentence false locally on the adjective and are also still risky in the sense that our test sentences may turn out to be true depending on the specific continuation (e.g. ... *that are inside of the circle*). The situation is different under the two expectation-based accounts. Firstly, under PRAGMATIC EXPECTATIONS, predictions differ between the newly added contexts as compared to the original complex contexts in Fig. 1c/d. This is because in the former, there is only one true continuation among the presented alternatives. In this sense, these types of complex contexts in Fig. 1e/i are comparable to the simple contexts in Fig. 1a/b. Therefore, the prediction from PRAGMATIC EXPECTATIONS is that the two newly added contexts should lead to faster RTs than the complex contexts in Exp. 1 because the adjective (e.g. *blue*) is more expected by virtue of allowing for true relative clause continuations. Secondly, under PRIMING, predictions depend on how visual salience is operationalized. If we think of salience of a color in terms of the fraction of the total area it covers, then predictions remain unchanged. If we instead think of it in terms of number of different colors shown in the picture, salience of the mentioned color (e.g. *blue*) will, by contrast, be altered, and presumably reduced, in the two new contexts because there is a third color beside the original two.

Moreover, we also have to consider the complexity effect in uni- vs. bicolored contexts in Exp. 1. This complicates predictions because at the moment we are lacking a theoretical understanding of this effect. Therefore, we do not know how it will be affected by the additional color in the newly added contexts and whether interacts with other factors affecting processing difficulty.

3.1. METHODS. The present experiment used the same general method and procedure as Exp. 1. The experimental design and materials were adjusted slightly by including additional contexts. The materials from Exp. 1 were reused and combined with the two new types of contexts exemplified in Fig. 1e/i. This resulted in twelve conditions in a 6×2 design with the factors CONTEXT (e.g. Fig. 1a–e,i) and PREPOSITION (*in-* vs. *outside*). In total, eight items per condition were constructed by combining the same colors and geometrical shapes as in Exp. 1, yielding 96 experimental sentences that were matched with 96 short filler sentences (each presented twice). Combined with 32 catch trials of the same type as in Exp. 1, this resulted in 224 trials that were distributed over four blocks using a Latin square design. Fifty native speakers of German who hadn't participated in Exp. 1 were recruited over the platform `prolific.co` and were paid for their participation. [2] The same exclusion criteria as in Exp. 1 resulted in the exclusion of four participants.

3.2. RESULTS. Accuracy across conditions is shown in Fig 3b. As in Exp. 1 participants' responses were correct in the majority of cases ($92.3 - 97.7\%$). Accuracy was lower and showed more variation in the bi- and tricolored as compared to the unicolored contexts. The logit mixed-

[2] There were 58 participants originally, but 8 of them had to be excluded because of an error with data transfer to the experimental server resulting in RT that could not be unambiguosly assigned to ROI.



(a) Mean RT on adjective in Exp. 2

(b) Accuracy in Exp. 2

Figure 3: Results from Exp. 2

effects model analysis (see section 2.1.1) revealed a marginally significant interaction between the two factors ($\chi^2(5) = 9.65, p < .086$; based on model comparison) showing that, in the two newly added conditions, the position of the homogeneous, i.e unicolored, color set had numerically opposite effects depending on the preposition in the sentence. When the homogeneous set was inside the container shape, the preposition *outside* led to slightly higher accuracy. By contrast, when the homogeneous set was outside the container shape, it led to slightly lower accuracy. In separate analyses of the two new contexts, we found that the interaction between CONTEXT and PREPOSITION was significant ($z = 2.27, p < .023$). However, the differences between prepositions was not reliable for either position of the homogeneous set (match inside: $z = -0.29, p = .766$; match outside: $z = 1.69, p = .091$). Within the other contexts, no effects were significant (all $|z| < 1.6$).

RTs on the adjective are shown in Fig. 3a. The general pattern in the uni- and bicolored contexts replicates the results from Exp. 1. Moreover, tricolored contexts had comparable RTs to bicolored contexts. In order to test our hypotheses, we computed a mixed model including the five contrasts in the second column of Table 1 for the six-level factor CONTEXT. Based on Exp. 1, we predicted that unicolored and locally true contexts would lead to shorter RTs than complex (bi- and tricolored) and locally false contexts, respectively (contrasts 1 & 2 in Table 1). Both predictions were borne out (simple vs. complex: $t = 6.75, p < .001$; locally true vs. false, one-tailed: $t = 1.81, p = .038$). For the remaining three contrasts, no directed prediction could be derived from our hypotheses, especially due to uncertainty regarding the contribution of the complexity effect, and thus we computed two-tailed hypothesis tests. We found that adjectives were read marginally faster after tricolored vs. bicolored contexts (contrast 3: $t = 1.93, p = .059$), but the location of the homogeneous color set did not affect RT significantly ($t = 0.93$, in bicolored contexts; $t = 1.58$, in tricolored contexts; contrasts 4 and 5, respectively).

Turning to the preposition, model comparisons revealed that, comparable to the results at the adjective, the preposition was read faster after unicolored than after bicolored contexts ($t = 6.52, p < .001$). Furthermore, the unicolored false contexts were read faster than the unicolored true ones ($t = -3.57, p < .001$). No other effects were significant on the preposition.

3.3. DISCUSSION. Although the newly added contexts, containing one object in a different color, were presumably at least as complex to remember as the bicolored contexts from Exp. 1, they did not lead to an increased complexity effect in RT and were even read faster on the color adjective, leading to marginal effect in the opposite direction. Assuming that tricolored contexts were more taxing on memory than bicolored ones, RT on the adjective thus seem to reflect a speed-up in anticipatory processes after tricolored contexts. Under PRAGMATIC EXPECTATIONS, such a speed-up in anticipatory processing could be due to the fact that only one of the possible continuations may yield a true output, e.g., *blue that are inside the circle* and is thus expected more strongly in tri- vs. bicolored contexts. At the same time, remembering three colors and their positions is more taxing on memory than remembering only two. Thus, total RT, reflecting both anticipatory and memory-related processes, may end up in a range similar to that of the bicolored contexts.

Under PRIMING, anticipatory processes on the color adjective may have been sped-up because of an increased salience of the relevant color term in the tricolored contexts. In these contexts, the mentioned color not only had the most representatives but also formed the only homogeneous set, i.e. covered all the object in its region. This may by itself induce an expectation that this is the color that will be referred to, again increasing expectations and resulting in a relative speed-up that may possibly be overlayed with an increased complexity effect. A salience-based explanation along these lines could, however, be questioned on the basis of the accuracy data. If the interaction between PREPOSITION and CONTEXT in the tricolored contexts is real, it would point to the additional object directing attention away rather than towards the homogeneous set. If this was in fact the case, increased salience of the color of the homogeneous set would be implausible.

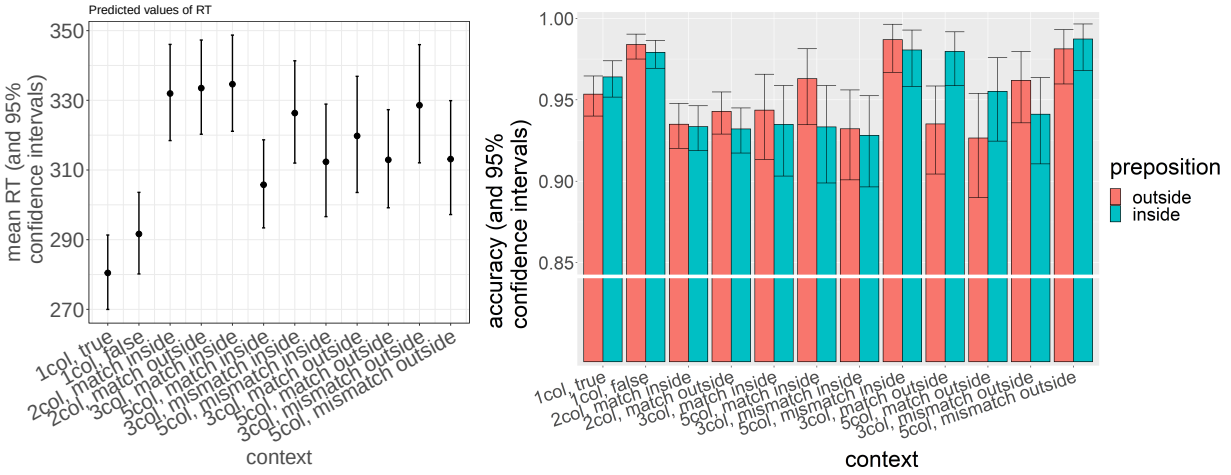
Finally, under REVISION SENSITIVITY we would have expected no differences in RT between multicolored contexts, apart from effects related to visual complexity and memory. This is because, in all these conditions, the sentence is false locally on the color adjective, but risky in the sense that it may still end up true given the right continuation. From this perspective the newly added, tricolored contexts were expected to behave exactly as the bicolored. Combined with their increased visual complexity we would have expected an increase rather than decrease in reading times on the adjective.

4. Experiment 3: Contrasting expectation- and memory-based effects. To disentangle effects that are based on memory demands from effects based on anticipatory processing as best as possible, we included an additional set of visual contexts in the present experiment which consisted of variations of the tricolored contexts in Exp. 2. Some of these contexts showed even more colors in the heterogeneous set. According to PRIMING, additional colors should affect the salience of color terms and lead to a decrease in RTs. These additional colors are, however, not expected to affect PRAGMATIC EXPECTATIONS beyond what salience contributes. This is because we assumed PRAGMATIC EXPECTATIONS to be driven by an expectation of true utterances (or utterances that match the context). Moreover, from the perspective of REVISION SENSITIVITY adding more colors to the heterogeneous set in contexts such as Fig. 1e/i does not change anything. Sentences in these conditions are still locally true but risky. These new contexts should, therefore, be processed just as the multicolored contexts in Exps. 1 & 2, modulo potential effects of visual complexity. In addition, we created a set of contexts in which we also changed the color of the homogeneous set (e.g. from blue to red, i.e. from matching to mismatching). This is expected to lead to a slowdown

under both PRIMING and PRAGMATIC EXPECTATIONS, either because of the increased salience of the mismatching color or because of the unexpectedness of an adjective that does not allow for true continuations (in the present design). Regarding memory demands, we would expect that contexts with more colors pose higher burdens on memory and thus lead to longer RT. The reasoning behind this expectation is that the more colors are shown on the picture the more information there is to be remembered, in particular the color of objects and their positions.

4.1. METHODS. The same materials as in Exp. 2 were used and these were combined with six further types of multicolored contexts, exemplified in Fig. 1f–h,j–l. Within the eight multicolored contexts that were not included in Exp. 1 (Fig. 1e–1l), three factors were manipulated in a 2 (NUMBER OF COLORS: 3 vs. 5) $\times$ 2 (HOMOGENEOUS COLOR: *match* vs *mismatch*) $\times$ 2 (POSITION OF HOMOGENEUOUS SET: *inside* vs. *outside*) factorial design. In combination with the original four contexts, this resulted in 24 conditions in a 12 (CONTEXT: see Fig. 1a–l) $\times$ 2 (PREPOSITION: *inside* vs. *outside*) within-items design. The entire set of visual contexts allows for a number of comparisons relevant to our three hypotheses and to the nature of the complexity effect from Exps. 1 & 2. These comparisons are described below in the results section. To keep the number of trials per participant in an acceptable range, especially for an online experiment, conditions were distributed between participants in eight lists. In each list, the four contexts from Exp. 1 were combined with two of the tri- or quintcolored contexts. Each context condition was included in two of the eight lists and that tri- and quintcolored contexts were never combined in one list. Context conditions were distributed across lists according to the following scheme (using labels from Fig. 1 to refer to conditions: List 1: e & i; List 2: f & j; List 3: f & l; List 4: h & l; List 5: h & j; List 6: e & k; List 7: g & k; List 8: g & i). The lists were generated completely analogously to Exp 2. and one of them was, in fact, identical to the materials in Exp. 2. In total, 281 native speakers of German (between 34 and 37 per list) who hadn't participated in one of the previous experiments or in any of the other lists were recruited over the platform `prolific.co` and were paid for their participation. The same exclusion criteria as before were used and resulted in exclusion of 53 participants in total from further analysis.

4.2. RESULTS. While accuracy was generally high ($92.6\% - 98.7\%$, see Fig. 4b), the highest values were observed in quintcolored contexts. For simplicity, we restricted the inferential statistics of accuracy to the tri- and quintcolored contexts. A logit mixed effects model analysis with fixed effects of NUMBER OF COLORS, HOMOGENEOUS COLOR, POSITION OF HOMOGENEOUS SET and PREPOSITION as well as random intercepts of participants revealed two significant interactions (POSITION OF HOMOGENEOUS SET $\times$ PREPOSITION: $z = 2.4, p = .019$; and NUMBER OF COLORS$\times$ HOMOGENEOUS COLOR: $z = -4.18, p < .001$). We resolved these interactions in separate analyses of tri- and quintcolored contexts. In fact, the pattern differed markedly between these two subsets: In the tricolored contexts, the general pattern from Exp. 2 was replicated. Thus, the preposition *outside* led to fewer errors if the homogeneous set was inside and vice versa. Although this pattern was descriptively visible in each of the tricolored contexts, it was only significant in conditions with mismatching contexts and preposition *inside* ($z = 2.69, p = .007$), leading to a significant three-way interaction in the tricolored contexts ($z = 2.45, p = .015$). By contrast, only a main effect of HOMOGENEOUS COLOR ($z = -4.89, p < .001$) was found in the quintcolored contexts because mismatching colors led to more correct responses than matching ones.



(a) Predicted RT at the adjective

(b) Accurcy in Exp. 3

Figure 4: Results from Exp. 3

The between-subjects design led to considerable variation in RT within the tri- and quintcolored contexts. To adjust for differences between participants, we provide a plot of predicted RT (i.e. marginal effects) from the mixed model analysis (Fig 4a). The mixed-effects model of RT at the adjective (see section 2.1.1) included comparisons analogous to those in Exps. 1 & 2 (i.e., contrasts 1–5 in the third column of Table 1) in addition to six further comparisons (contrasts 6–11 in Table 1) that are restricted to tri- and quintcolored contexts and tested for specific predictions of PRIMING and PRAGMATIC EXPECTATIONS. The contrasts 6/9 tested for a prediction of PRAGMATIC EXPECTATIONS (in contexts with the homogeneous set inside or outside the container shape, respectively) that adjectives like *blue* should be read faster if they start the only true continuation (e.g. 1e/f) than if a different adjective starts the only such continuation (e.g. 1g/h). The contrasts in 7/10 and 8/11 tested for the prediction of PRIMING (with the homogeneous set inside or outside, respectively) that a relatively large number of objects in a distracting color in a matching contexts (e.g. Fig 1e vs.f and 1i vs. j) or a relatively small number of matching objects in a mismatching contexts (e.g. Fig 1g vs. h and 1k vs. l) is expected to cause slow down.

The analysis revealed significant effects of the contrasts 1, indicating shorter RT in unicolored contexts in comparison to the mean RT of all the others ($t = 21.63, p < .001$), 2, indicating shorter RT for true vs. false unicolored contexts ($t = 4.57, p < .001$), and 3, showing longer RT after bicolored than after tri- and quintcolored contexts ($t = 7.286, p < .001$). These three effects were replicated from Exp. 2 (and partly Exp. 1). Moreover, contrast 8, comparing the number of mismatching (e.g. red) objects in *three* vs. *five colors, match inside* ($t = -5.01, p < .001$), and contrast 10, comparing the number of matching objects in *three* vs. *five colors, mismatch outside* (one-tailed: $t = -2.72, p = .003$), were significant. And finally, the contrast 7 was significant ($t = -2.48, p = .013$) showing that more objects in a matching color led to a slow down in *mismatch inside* contexts. None of the other contrast were significant ($|t| < 1.3$). Looking at the direction of the effects, only contrast 8 seems to speak to one of our hypotheses, namely PRIMING, because it seems to show that a larger number of objects in a mismatching color slows

down reading. However, all of the effects in the contexts with three and five colors could also be explained by a general trend for faster RT in contexts with five vs. three colors in addition to an interaction between HOMOGENEOUS COLOR and POSITION OF HOMOGENEOUS SET in the tricolored contexts. To test for this explanation, we opted for a post-hoc, factorial analysis of the tri- and quintcolored contexts (cf. description of factors above). This analysis revealed a three-way interaction in ($t = 2.52, p = .012$), which we resolved by conducting separate analysis of the tri- and quintcolored contexts. There was a two-way interaction in the former ($t = -2.52, p = .012$) but not the latter ($t = 0.17, p = .87$). The interaction in the tricolored contexts was due to a marginally significant difference between the two *matching* ($t = -1.68, p = .094$) and an opposite but non-significant difference in the *mismatching* contexts ($t = 1.09, p = .276$).

On the preposition, we observed an intricate pattern of RT. Because there was no truth-value effect, we refrain from reporting it in detail. One finding that seems worth mentioning, however, was a general tendency for longer RT in matching sentences (i.e., sentences that were potentially true at the adjective) vs. mismatching sentences. In multicolored contexts with the HOMOGENEOUS SET inside the container shape, this was the only significant effect ($t = 2.8, p = .005$). When the HOMOGENEOUS SET was outside, it interacted with the NUMBER OF COLORS but was significant within both levels of this factor (3 cols.: $t = 4.24, p < .001$; 5 cols.: $t = 8.9, p < .001$).

4.3. DISCUSSION. While PRAGMATIC EXPECTATIONS can provide a plausible explanation of the results from Exps. 1 & 2 (at least in combination with certain assumptions about memory-related effects), its predictions were disconfirmed in the present experiment. In particular, RTs did not differ between contexts that allowed for true and false continuations. Moreover, PRIMING only correctly predicted the difference between contexts with three and five colors. In general, contexts with more colors actually led to shorter RT, contrary to our initial expectations. One possible explanation could be that the more colors that were present, the better participants were able to focus on the homogeneuos set and ignore the distractor objects. This would actually be a useful adaptation to the present design, because only the homogeneous sets were relevant for the task.

Taking into account this reversed complexity effect, the data on the adjective seem compatible with REVISION SENSITIVITY. However, we did not find an effect of truth on the preposition. Instead, we saw that the conditions that were potentially true on the color adjective led to prolonged RTs on the preposition, regardless of their actual truth values at that point. We take this effect as an indication of memory retrieval of the position of the matching color set. If this interpretation is on the right track, our entire data set would call for an explanation in which truth-evaluation interacts in intricate ways with effects that are related to the retention of information in and retrieval of that information from memory. This interpretation also seems to be consistent with the interactions we observed in the tricolored contexts in RT on the adjective and also in accuracy. These interactions indicate that the one objects that had a different color in these conditions may modulate memory load by attracting attention to varying degrees depending on the position of the homogeneous set.

5. General Discussion. Overall, the present experiments show that self-paced reading is affected by the relation between a previously presented visual context and the incremental processing of the compositional meaning of a quantified sentence. In particular, reading times were sensitive towards a match between context and sentence meaning (e.g. the truth-value related effects in Exp. 1) and also towards interactions between picture complexity and task demands (e.g. the difference

between uni- and multicolored pictures and its modulation across the three experiments). We also observed RT effects that seemed to depend on a combination of these two factors (e.g. the effect of matching vs. mismatching colors at the preposition in Exp. 3).

Crucially, none of our three hypotheses can account for our entire data set. First of all, REVISION SENSITIVITY predicted truth-value effects as soon as they can be unambiguously decided, but, on the preposition, such effects were only observed in Exp. 1. Secondly, PRAGMATIC EXPECTATIONS predicted RT effects already on the adjective, which is the position where the potential for true continuations is decided. This was directly tested in Exp. 3 but not confirmed. Finally, PRIMING predicted that the number of objects that match or mismatch the color adjective should affect RTs, irrespective of the (potential) truth value, which was also not confirmed in Exp. 3.

In the following, we suggest two possible explanations for this discrepancy between predictions and results. Firstly, some of the predicted effects could have remained undetected because of insufficient power. In particular, we consider it possible that the complexity effects may have obscured a potential truth-value effect. This appears likely if we consider the relatively large magnitude (and associated variance) of the complexity effect compared to the truth-value effect that we consistently observed within the unicolored contexts. However, note that we did not have any specific prior expectations about these effects. Another, factor that may have limited the chance to detect truth-value related effects could be participant variance caused by the between-design in Exp. 3. This points to a general methodological challenge in distinguishing between the three tested hypotheses: Because of the close relation between these hypotheses, distinguishing between them requires specific comparisons and the resulting number of relevant conditions may exceeded the limits of within-designs. One way to approach this issue would be to use continuous independent variables to manipulate PRAGMATIC EXPECTATIONS and PRIMING in a gradual fashion.

Secondly, the discrepancy between predictions and results may also be related to working memory limitations. Specifically, the task of encoding all information in the context pictures and using this information to build expectations about all possible continuations in our task is cognitively highly demanding. Indeed, there are theoretical considerations on effects of restricting the memory capacity underlying pragmatic interpretation (Franke et al. 2011) as well as proposals that make expectations during sentence processing contingent on imperfect memory representations (e.g. Futrell et al. 2020) or on adaptations to the experimental context (see e.g. Schlotterbeck et al. 2022; for an application to quantifier processing).

If the observed discrepancy is due to the second, theoretically more interesting explanation, we would still be missing crucial pieces to the puzzle of incremental compositional interpretation. We think that applying refined notions of the relation between memory and expectations to semantic and pragmatic processing may lead us a crucial step forward towards finding missing pieces. Moreover, such considerations can be combined with assumptions about adaptive processes as well as specific theories about the representations and processes involved in compositional interpretation (see e.g. Bott et al. 2019, Bremnes et al. 2022, Pietroski et al. 2009; for application to quantifiers) in order to refine the type of hypotheses tested in the present study. If such hypotheses are tested using a mix of different methods, we may obtain a better understanding of the processing correlates of compositional interpretation.

References.

Augurzky, P., O. Bott, W. Sternefeld & R. Ulrich. 2017. Are all the triangles blue?–ERP evidence for the incremental processing of German quantifier restriction. *Language and Cognition* 9(4). 603–636. https://doi.org/10.1016/j.cognition.2017.05.023.

Augurzky, Petra, Michael Franke & Rolf Ulrich. 2019. Gricean expectations in online sentence comprehension: An ERP study on the processing of scalar inferences. *Cognitive Sci* 43. https://doi.org/10.1111/cogs.12776.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01.

Bott, Oliver, Fabian Schlotterbeck & Udo Klein. 2019. Empty-Set Effects in Quantifier Interpretation. *Journal of Semantics* 36(1). 99–163. https://doi.org/10.1093/jos/ffy015.

Bott, Oliver & Wolfgang Sternefeld. 2017. An Event Semantics with Continuations for Incremental Interpretation. *Journal of Semantics* 34(2). 201–236. https://doi.org/10.1093/jos/ffw013.

Bremnes, H.S., J. Szymanik & G. Baggio. 2022. Computational complexity explains neural differences in quantifier verification. *Cognition* https://doi.org/10.1016/j.cognition.2022.105013.

Frank, M. C. & N. D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998–998. https://doi.org/10.1126/science.1218633.

Franke, Michael, Gerhard Jäger & Robert van Rooij. 2011. Vagueness, signaling and bounded rationality. In Takashi Onada, Daisuke Bekki & Elin McCready (eds.), *New frontiers in artificial intelligence*, 45–59. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-25655-4_5.

Futrell, Richard, Edward Gibson & Roger P. Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science* 44(3). https://doi.org/10.1111/cogs.12814.

Pietroski, Paul, Jeffrey Lidz, Tim Hunter & Justin Halberda. 2009. The meaning of 'most': Semantics, numerosity, and psychology. *Mind and Language* 24(5). 554–585. https://doi.org/10.1111/j.1468-0017.2009.01374.x.

Schlotterbeck, Fabian, Petra Augurzky & Rolf Ulrich. 2022. Degree of incrementality is modulated by experimental context – ERP evidence from German quantifier restriction. *Language, Cognition and Neuroscience* https://doi.org/10.1080/23273798.2022.2125541.

Szymanik, Jakub. 2016. *Quantifiers and Cognition. Logical and Computational Perspectives*. Springer. http://doi.org/10.1007/978-3-319-28749-2.

van Tiel, Bob, Michael Franke & Uli Sauerland. 2021. Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences of the United States of America* 118(9). https://doi.org/10.1073/pnas.2005453118.