Processing particularized pragmatic inferences under load

Margarita Ryzhova (mryzhova@coli.uni-saarland.de)
Saarland University, Campus C7.2, 66123 Saarbrücken, Germany

Vera Demberg (vera@coli.uni-saarland.de)
Saarland University, Campus C7.2, 66123 Saarbrücken, Germany

Abstract

A long-standing question in language understanding is whether pragmatic inferences are effortful or whether they happen seamlessly without measurable cognitive effort. We here measure the strength of particularized pragmatic inferences in a setting with high vs. low cognitive load. Cognitive load is induced by a secondary dot tracking task. If this type of pragmatic inference comes at no cognitive processing cost, inferences should be similarly strong in both the high and the low load condition. If they are effortful, we expect a smaller effect size in the dual tasking condition. Our results show that participants who have difficulty in dual tasking (as evidenced by incorrect answers to comprehension questions) exhibit a smaller pragmatic effect when they were distracted with a secondary task in comparison to the single task condition. This finding supports the idea that pragmatic inferences are effortful.

Keywords: experimental pragmatics; redundancy; cognitive costs; dual-tasking

Introduction

Language understanding involves recovering the intended meaning of the speaker, which often goes far beyond the literal semantic meaning of a discourse. In the case of pragmatic inferences, listeners must access and integrate lots of additional relevant information such as situational and linguistic context, world knowledge, or speaker personality. A long-standing open question in pragmatics is to what extent this process of executing pragmatic inferences is cognitively demanding. The existing accounts vary substantially in their views on this point: the Default Model states that pragmatic inferences arise by default and are not associated with cognitive costs (Levinson, 2000; Chierchia et al., 2004). The equally influential Contextual Hypothesis originates in the Relevance theory and claims that pragmatic inferences are only generated in relevant contexts and might be associated with processing difficulty (Wilson & Sperber, 2012; Carston, 1998; Degen & Tanenhaus, 2019).

Recent studies in experimental pragmatics have not been able to conclusively resolve this question: While some studies find evidence for effects of processing difficulty (Bott & Noveck, 2004; De Neys & Schaeken, 2007; Dieussaert, Verkerk, Gillard, & Schaeken, 2011) related to pragmatic inferences, methodological criticisms have been voiced regarding of some of these studies (see Zondervan (2010) for a discussion), and other studies report that no costs of pragmatic inferences could be measured, thus supporting the Default model (Feeney, Scrafton, Duckworth, & Handley, 2004; Grodner, Klein, Carbary, & Tanenhaus, 2010; Marty, Chemla, & Spector, 2013). We note that most of the studies so far have investigated the question of processing effort related to pragmatic inferences for scalar implicatures, which are a type of generalized pragmatic implicatures.

In the present study, we investigate cognitive cost associated with inferring particularized pragmatic implicatures, triggered by informationally redundant utterances. The pragmatic inferences triggered by informationally redundant utterances are more context-dependent, and might hence be more prone to involve measurable cognitive effort.

The notion of informational redundancy (IR) refers to materials which are easily predictable from listeners' world knowledge and based on the conversational context. For example, in the following passage, the utterance in bold is redundant since it can be conventionally inferred based on its precedent.

Lisa went swimming. ***She brought her swimsuit!***

Kravtchenko and Demberg (2015) showed that when people encounter such informationally redundant utterances, an inference that Lisa usually forgets her swimsuit may be triggered. That is, comprehenders alter their beliefs about activity typicality (*bringing a swimsuit*) and rated the probability of Lisa usually bringing her swimsuit lower when the IR utterance was mentioned in comparison when it was not.

In the present study, we took the materials from (Kravtchenko & Demberg, 2015) and used a dual-task paradigm to manipulate the amount of available cognitive resources. We expect to observe that the secondary task reduces the amount of cognitive resources available for the language comprehension task, and thus affects participants' likelihood of drawing pragmatic inferences, if these inferences are cognitively effortful. Alternatively, it could be that the likelihood of pragmatic inferences is unchanged, but performance on the secondary task is reduced while the pragmatic inference is drawn. This would also indicate that pragmatic inferences are effortful.

If, on the other hand, pragmatic inferences are effortless, we should observe no reduction in performance on either the secondary tracking task nor on the likelihood of drawing the inferences.

Processing cost of pragmatic inferences

Different experimental designs have been proposed in the literature to test for the cognitive cost of inferring pragmatic implicatures. De Neys and Schaeken (2007) used a dual task design including a spatial memory task and a truth judgment task on scalar implicatures in underinformative sentences such as *Some tuna are fish*. They assessed recall performance on the spatial memory task as a measure of cognitive load. In their design, participants first had to memorize a pattern of three dots that appeared on the 3x3 screen for a short amount of time, before being asked to judge the truth of a sentence. They then had to reproduce the dot pattern. De Neys and Schaeken (2007) report a significantly decreased rate of pragmatic responses in the high load condition, where the dots were randomly distributed across the grid, compared to the low load condition (the dots were grouped along vertical or horizontal axes). Additionally, pragmatic responses in the dual task condition were 700 ms slower than under the single task. These results indicate that scalar implicatures require processing effort.

Bott and Noveck (2004) use time pressure to detect cognitive load related to the inference of scalar implicatures. In their experiment 4, participants had a limited time to judge the truth of underinformative sentences (900 ms in high load condition vs. 3000 ms in low load condition). The number of pragmatic responses was significantly lower when participants were forced to answer more quickly. Moreover, in their experiment 3 (which did not include any load manipulation), Bott and Noveck (2004) found that those participants who answered pragmatically took significantly longer to respond, in comparison with those who provided literal answers. Similarly, Huang and Snedeker (2009) demonstrated that referent identification in underinformative sentences with *some* was significantly delayed relative to non-underinformative sentences, and concluded that these implicatures are costly.

These conclusions are however controversial: Grodner et al. (2010) found different results to those in (Huang & Snedeker, 2009) and proposed that the observed delay in pragmatic responses might be associated with additional time needed to integrate the interpretation with the context rather than with the processing difficulty of the implicature itself.

Marty et al. (2013) observed mixed evidence on the cost of pragmatic inferences using a dual task paradigm. For cognitive load manipulation, participants were told to memorize the sequence of letters before the main task (four letters in the high load condition vs. two letters in the low load). The main task consisted of a sentence-picture verification task including two types of scalar implicatures (underinformative sentences with a quantifier *some* and numerals (*4 dots are red*)). While there were significantly less pragmatic responses under high cognitive load for underinformative sentences with *some*, no effect of load was found on the stimuli including numerals.

Numerals can be argued to be the strongest form of generalized pragmatic implicatures, which are strongly trained, and like scalar implicatures, largely independent of context. We here instead study particularized implicatures, which require the comprehender to integrate the linguistic signal with situational knowledge and world knowledge. We hypothesize that investigating such types of implicatures may shed additional light on the question of costliness of pragmatic inferences.

Method

Participants

382 eligible participants (*mean age* = 34 yrs; 60% female) were recruited via the crowdsourcing platform Prolific. The task was open only to workers who stated English as their native language, and who had an approval rating of $> 95\%$. All participants reported no hearing problems and had normal or corrected-to-normal vision.

Procedure

Language Task The language comprehension task consisted of listening to four short stories. The stories were adapted from (Kravtchenko & Demberg, 2015) and read out by a native speaker of American English. A story consisted of the context stating the topic (e.g., *grocery shopping*, *going swimming*) and an introduction of the story characters (2-3 characters per story). The critical sentence which gives rise to the pragmatic inference consists of a highly predictable activity in the context of the scenario (e.g., *paying the cashier*, *bringing a swimsuit*) (see Table 1, a). We call the mention of the predictable activity the "informationally redundant" (IR) utterance. A typical item can be seen in Table 1, a and b. The informationally redundant utterance was recorded with exclamatory intonation.

The without-IR story condition consists only of the context (part a in Table 1). The with-IR condition consists of both the context and the informationally redundant utterance (Table 1, a and b).

Table 1: Example of the "Going swimming" story

a. Context
Lisa likes to go swimming at a nearby pool after work. A couple days ago she was at the pool when she saw Harvey, another regular member, and they stopped to chat. After Harvey changed and went out into the pool area, he ran into Jen, another swimmer and a friend of Lisa's.
b. Optionally mentioned IR activity description (in bold)
Harvey said to Jen: "Lisa's here to swim, too. **She brought her swimsuit!**"

Participants were instructed to listen to the stories carefully and answer three story-related questions, which appeared on separate screens. One of the questions was aimed to assess participants' judgments about the typicality of the informationally-redundant activity (target question: *How often do you think Lisa usually brings her swimsuit, when going*

swimming?). A second question addressed an activity that is generally non-predictable from the script (filler question: *How often do you think Lisa usually brings her children, when going swimming?*). To answer these questions, participants could indicate their estimates using a slider that ranged from 0 ('*Never*') to 100 ('*Always*'). The order of the questions was randomized. After answering the target and filler questions, participants were shown the third question about the content of the story (comprehension question: *What does Lisa like to do after work?*). The question was used to check whether participants listened to the stories carefully.

To avoid stereotypical responses, each participant also saw four filler stories without IR manipulation but of a similar structure. In total, we had 20 different story topics, each presenting in 2 story conditions. All items were randomized to ensure that each participant encountered each condition only once, including the story topic.

Dual Task In order to manipulate the amount of available cognitive resources, we used a dot tracking task, which is available from the website of Cognition Laboratory Experiments, designed by John H. Krantz[1]. We used the dot-tracking task as an easy-to-run-online analog of the ConTRe task (Mahr, Feld, Moniri, & Math, 2012) for measuring the effects of a workload on a continuous course of a task with high precision.

In half of the trials, participants listened to the stories in parallel with following the dot, which randomly moved on the screen (high load condition). In the other half of the trials, they performed only listening and, instead of the dot, saw the cross in the middle of the screen (low load condition).

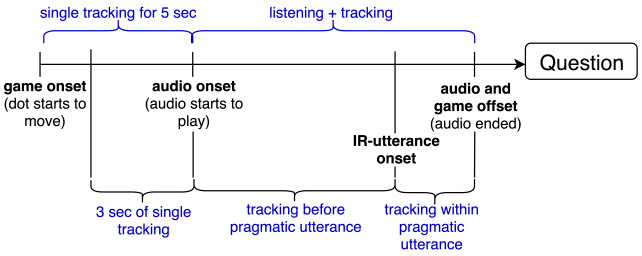

Figure 1: The timecourse of a trial in the high load condition

Each new trial in the high load condition started with the dot appearing in the middle of the screen. The dot began to move only after the participant hovered the cursor to the dot. They were instructed to follow the dot carefully with their mouse throughout the whole trial and keep the cursor as close to the dot as possible. After the dot started moving, participants had 5 seconds of single-tracking before the audio began to play. Once the story ended, participants were redirected to the page with judgment questions. For the analysis, we also annotated the onset of the pragmatic utterance (*She brought her swimsuit!*) - see Figure 1 for the time course of one trial in the high load condition.

The dot was controlled with three parameters: maximum angle variation, speed, and size of the dot. Based on the preliminary testing, we balanced the parameters such that tracking required a significant amount of cognitive resources but would not dominate the listening[2]. The sampling rate for the dot and the cursor coordinates was set to 20 Hz.

Results

All results were analysed using linear mixed effects models, as implemented in the lme4 library (Bates, Mächler, Bolker, & Walker, 2015) in R. The linguistic task used participants' ratings in target question as a response variable, while the dot-tracking task was analysed using tracking deviations as a response variable. P-values were obtained using the Satterthwaite approximation for degrees of freedom, as implemented in the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017).

For the model of linguistic judgments, as fixed effects, we included story (with-IR vs. without-IR story), load (high vs. low cognitive load), and subject group (subjects who answered all comprehension questions correctly vs. those who made at least one mistake).

For models of dot-tracking deviation, we calculated the deviations as the Euclidean distance between the dot and the cursor in each timestamp. Fixed effects for these models included the same set of predictors. All factors were sum coded.

We always started out by fitting models with the maximal random effects structure justified by the design. Thus, for ratings' models we included by-subject random intercepts and slopes for story and load conditions as well as by-item random intercepts and slopes for both factors and their interaction. By-subject random slopes for the interaction were not included in the model, because we did not have any repeated measures for the interaction (each subject saw each condition only once). Since in this experiment, we had relatively few data points per subject, models with by-subject random effects did not always converge. In the case of non-convergence, we simplified the random effect structure progressively until convergence was achieved (Barr, Levy, Scheepers, & Tily, 2013), see model specifications below.

Language task

The main analysis included the analysis of ratings given to the target question (see Table 2).

The model of linguistic judgments showed a significant main effect of informational redundancy (see Table 2): if the utterance was included in the story (with-IR condition), participants' typicality ratings were significantly lower than when it was not mentioned (see Figure 2). In line with Kravtchenko and Demberg (2015), this finding suggests that

[1] https://psych.hanover.edu/JavaTest/CLE/Cognition_js/exp/dualTask.html

[2] size of the dot = 30, dot speed = 600, maximum angle variation = 180

participants make pragmatic inferences and accommodate redundancy by lowering their beliefs about otherwise highly predictable activities.

The effect of Story-Load interaction was not statistically significant, suggesting that pragmatic processing is not influenced by a burden placed on the participants. However, as it was shown in Dieussaert et al. (2011), participants may respond differently to dual task demands, depending on individual differences in working memory (also discussed in Feeney et al. (2004)) or other executive functions. An analysis of the comprehension questions revealed that a substantial number of questions was answered incorrectly, especially in the high load condition ($b = -0.32$, $SE = 0.16$, $z = -2.03$, $p < .05$ *). This reveals that some participants may have struggled in the dual task setting. In a post-hoc analysis, we therefore split up the data set according to participants' answering accuracy to the semantic question.

Table 2: Effect sizes (b), standard errors (SE), t-values, and p-values for the LMER model of linguistic judgements. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*t*	*p*
Intercept	82.49	1.68	49.09	***
Story: with IR	-6.37	1.57	-4.05	***
Load: high	0.81	0.93	0.87	ns
Story*Load	2.66	1.79	1.49	ns
Random Effects		*Variance*		
Subject		123.15		
story | Subject		187.80		
load | Subject		22.15		
Item		45.98		
story | Item		23.51		

We found that roughly half of the participants answered all four comprehension questions correctly (185 subjects; *mean age* = 35 yrs; 61% female), while the other half made one or more mistakes in answering the questions (197 subjects; *mean age* = 35; 60% female). For the following analyses, we derived from this result a grouping variable distinguishing between the participants who made mistakes vs. those that did not.

In the updated model of linguistic judgements (see Table 3), there was a significant main effect of subject group, which led us to analyze both groups of subjects separately. A model including only those participants who answered the comprehension questions correctly showed a significant main effect of a story, indicating that participants made pragmatic inferences but the secondary task did not influence their ratings (see Table 4).

In contrast, the model for participants who made at least one mistake in answering comprehension questions showed a significant interaction between story and cognitive load (see Table 5). A separate analysis of data split by high vs. low load condition showed that this interaction is driven by larger effect sizes of the IR-utterance when the load is low ($b = -10.87$, $SE = 2.33$, $t = -4.67$, $p < .001$ ***) than when it is high ($b = -5.52$, $SE = 2$, $t = -2.76$, $p < .01$ **) – see Figure 2. This means that for participants who had more trouble answering the comprehension questions, the pragmatic effect was on average smaller when they were distracted by a secondary task in comparison when they were not.

Table 3: Effect sizes (b), standard errors (SE), t-values, and p-values for the LMER model of linguistic judgements with subjects grouping. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*t*	*p*
Intercept	82.56	1.66	49.78	***
Story: with IR	-6.33	1.59	-3.98	***
Load: high	0.80	0.93	0.86	ns
Subject group: correct	4.46	1.44	3.09	**
Story*Load	2.60	1.79	1.45	ns
Story*Subject group	3.17	2.30	1.38	ns
Load*Subject group	-0.82	1.89	-0.43	ns
Story*Load*Subject group	-4.22	3.66	-1.15	ns
Random Effects		*Variance*		
Subject		118.52		
story | Subject		186.23		
load | Subject		22.49		
Item		44.77		
story | Item		24.53		

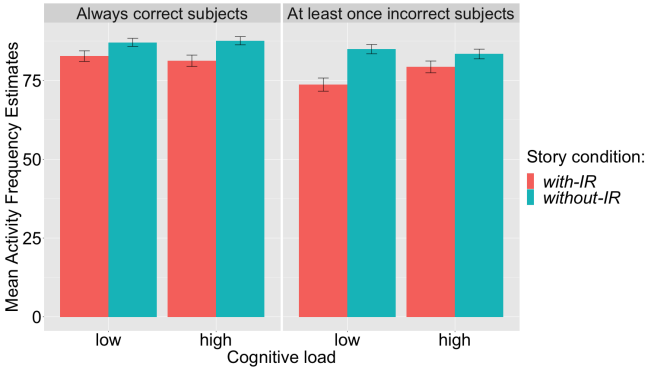

Figure 2: Mean participants' ratings (± SEM) in target pragmatic question in subjects who answered all comprehension questions correctly vs. those who made at least one mistake, depending on cognitive load and story conditions.

Dual task

In high load condition, participants performed listening in parallel with dot tracking. For the analysis of tracking deviations, we calculated by-subject mean tracking deviations in the single tracking interval (to reduce the noise, we excluded

Table 4: Subjects who answered all comprehension questions correctly. Effect sizes (b), standard errors (SE), t-values, and p-values for the LMER model of linguistic judgements. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*t*	*p*
Intercept	84.85	1.85	45.86	***
Story: with IR	-4.65	1.97	-2.36	*
Load: high	0.35	1.17	0.30	ns
Story*Load	0.63	2.35	0.27	ns
Random Effects		*Variance*		
Subject		93.20		
story | Subject		172.33		
Item		51.43		
story | Item		31.05		

Table 5: Subjects who made at least one mistake in comprehension questions. Effect sizes (b), standard errors (SE), t-values, and p-values for the LMER model of linguistic judgements. Significance codes: *** .001 | ** .01 | * .05

	b	*SE*	*t*	*p*
Intercept	80.25	1.84	43.58	***
Story: with IR	-7.82	1.93	-4.06	***
Load: high	1.08	1.47	0.73	ns
Story*Load	5.39	2.71	1.99	*
Random Effects		*Variance*		
Subject		147.67		
story | Subject		207.44		
load | Subject		64.47		
Item		43.69		
story | Item		16.69		

the first two seconds of tracking), tracking interval along the whole audio period, and in the tracking intervals before and after the onset of pragmatic utterance (see Figure 1).

First, we analyzed the effect of listening on dot-tracking. Thus, we compared the mean subjects' tracking deviations in the single tracking interval vs. the whole audio interval (dual interval). We built a linear regression mixed effects model with a log-transformed dependent variable (see Table 6), and also included the subject's group as a predictor. The model showed a significant main effect of the interval ($b = -0.07$, $SE = 0.02$, $t = -4.97$, $p < .001$ ***) suggesting that people performed less well in tracking when listening to language. There was also a marginally significant effect of the subject group ($b = -0.08$, $SE = 0.05$, $t = -1.72$, $p = 0.086$), showing that people who later on showed difficulty in answering a comprehension question generally had more difficulty with the dot-tracking task (see Figure 3).

Second, we compared the tracking interval before the onset of the pragmatic utterance with the tracking interval after the onset of pragmatic utterance. There were found, however, no differences in performances between the intervals. We hypothesized that pragmatic processing could happen shortly after participants met the material requiring pragmatic processing. Though, it is yet not clear when pragmatic processing actually happens. In fact, it might occur later after participants faced the target question.

Table 6: Effect sizes (b), standard errors (SE), t-values, and p-values for the LMER model of tracking deviations in single vs. dual intervals. The response variable was log-transformed. Significance codes: *** .001 | ** .01 | * .05 | . 0.1

	b	*SE*	*t*	*p*
Intercept	4.64	0.02	196.72	***
Interval: single	-0.07	0.02	-4.97	***
Subject group: correct	-0.08	0.05	-1.72	.
Interval*Subject group	-0.008	0.03	-0.27	ns
Random Effects		*Variance*		
Subject		0.19		

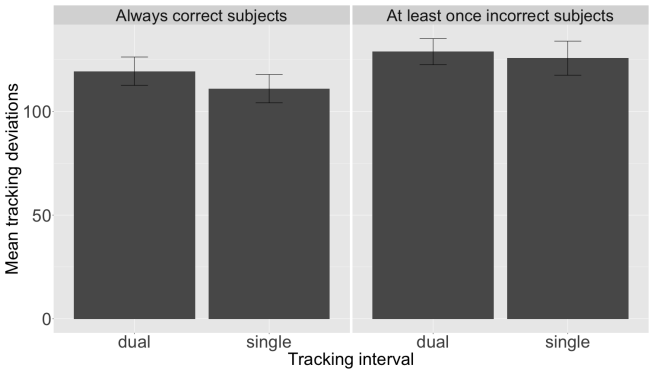

Figure 3: Mean tracking deviations (± SEM) in the single tracking interval (single) vs. the whole audio interval (dual) for two groups of participants.

Discussion

In the current literature, different accounts about the nature of pragmatic inferences exist, ranging from those that state automatic and seamless processing (Levinson, 2000) to those that advocate a cognitively demanding view of pragmatic processing. The latter view provides a possible explanation for why an intended message might sometimes not be decoded correctly by a cognitively overloaded addressee (Bott, Bailey, & Grodner, 2012). However, current proposals are mostly based on evidence from experiments involving only a small range of types of pragmatic implicatures, mostly scalar implicatures. Our study contributes processing evidence from particularized pragmatic inferences to the literature.

In this study, we investigated the role of cognitive workload in pragmatic processing associated with informationally

redundant utterances. Inferences triggered by informationally redundant utterances imply altered beliefs about activity typicality mentioned in a highly relevant context. Thus, in story materials, we manipulated the presence or absence of informationally redundant utterances describing activities that are anticipated from the story topic without an overt mentioning. By introducing a secondary dot-tracking task, we increased cognitive load, which we expected would affect comprehenders' likelihood to derive pragmatic inferences, if they are cognitively costly.

We replicated the overall finding reported in Kravtchenko and Demberg (2015) in auditory settings: participants' typicality ratings were significantly lower when the predictable activity was mentioned explicitly in the story, in comparison with when it was not. Hence, participants accommodated informational redundancy by lowering their beliefs about activity typicality.

A key finding of our study is a statistically significant interaction of informational redundancy and cognitive load among the group of participants who did less well in the comprehension questions. Under high cognitive load, subjects in this group showed a smaller effect of pragmatic inferencing than in the low load condition. On the other hand, the group of participants who showed no difficulty in the comprehension questions did not show any significant differences in pragmatic inferencing as a function of load condition. This finding is in line with an interpretation where participants who experience difficulty under high load show a reduced pragmatic inferencing effect.

Previous studies have also shown that whether a pragmatic inference is drawn may depend on a variety of listener characteristics such as age, working memory capacity, personal traits, or pragmatic skills (Dieussaert et al., 2011; Antoniou, Cummins, & Katsos, 2016; Katsos & Bishop, 2011; Noveck, Bianco, & Castry, 2001). In a study of scalar inferences, Dieussaert et al. (2011) found that participants with lower working memory span provided fewer pragmatic responses under high cognitive load. In contrast, the answers of participants with higher working memory capacity were not affected by increased cognitive load. Dieussaert et al. (2011) explain this finding from a contextualist point of view: if scalar implicatures are costly in the sense that they require WM capacity, listeners who have low working memory capacity may be less likely to derive a pragmatic inference when cognitive resources are loaded.

Our data however also show that, numerically, the pragmatic effect under low load was larger than the pragmatic effect observed in the group of always correct subjects (see Figure 2, the difference between mean ratings in with- vs. without-IR story conditions). This might be considered as contrary to the findings in (Dieussaert et al., 2011) where the low WM group, under low load condition, was not more pragmatic than the group with high memory capacity. Note though that this numerical difference is not statistically reliable, and must hence be interpreted with caution.

The studies by (Dieussaert et al., 2011) differ from ours in that they can specifically relate the differences in pragmatic effects to working memory, while our study used answers to comprehension questions to divide up the groups, a measure which may reflect attention or task switching more than working memory capacity.

The idea that participants who answered questions incorrectly may have more difficulty in cognitive control and task switching is also consistent with the observation that these participants had also more difficulty in the dot tracking task. Thus, it is crucial to relate individual differences in cognitive control: the group of participants who might have more difficulty in switching between the tasks (see e.g., Häuser, Demberg, and Kray (2018)) may have insufficient cognitive resources left for pragmatic inferences if these are cognitively costly.

In future work, the relation between task performance and various individual differences such as working memory span, multitasking abilities and linguistic experience should be investigated using separate measures of cognitive control (Lavie, 2010).

Taken together, these findings suggest that pragmatic processing is not entirely automatic and requires cognitive effort, consistent with the contextualist view.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. We thank the anonymous reviewers for their insightful comments.

References

Antoniou, K., Cummins, C., & Katsos, N. (2016). Why only some adults reject under-informative utterances. *Journal of Pragmatics*, *99*, 78–95.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bott, L., Bailey, T. M., & Grodner, D. (2012). 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 inferences. *Journal of memory and language*, *51*(3), 437–457.

Carston, R. (1998). Informativeness, relevance and scalar implicature. *Pragmatics And Beyond New Series*, 179–238.

Chierchia, G., et al. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and beyond*, *3*, 39–103.

Degen, J., & Tanenhaus, M. K. (2019). Constraint-based pragmatic processing. *Handbook of Experimental Semantics and Pragmatics*.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental psychology*, *54*(2), 128–133.

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: Further evidence that scalar implicatures are effortful. *The Quarterly Journal of Experimental Psychology*, *64*(12), 2352–2367.

Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *58*(2), 121.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*(1), 42–55.

Häuser, K. I., Demberg, V., & Kray, J. (2018). Surprisal modulates dual-task performance in older adults: Pupillometry shows age-related trade-offs in task performance and time-course of language processing. *Psychology and aging*, *33*(8), 1168.

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive psychology*, *58*(3), 376–415.

Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67–81.

Kravtchenko, E., & Demberg, V. (2015). Semantically underinformative utterances trigger pragmatic inferences. In *Cogsci.*

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *Journal of statistical software*, *82*(13).

Lavie, N. (2010). Attention, distraction, and cognitive control under load. *Current directions in psychological science*, *19*(3), 143–148.

Levinson, S. C. (2000). *Presumptive meanings*. na.

Mahr, A., Feld, M., Moniri, M. M., & Math, R. (2012). The contre (continuous tracking and reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity. In *Adjunct proceedings of the 4th international conference on automotive user interfaces and interactive vehicular applications* (pp. 88–91).

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

Noveck, I. A., Bianco, M., & Castry, A. (2001). The costs and benefits of metaphor. *Metaphor and Symbol*, *16*(1-2), 109–121.

Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. Cambridge University Press.

Zondervan, A. (2010). *Scalar implicatures or focus: an experimental approach*. Netherlands Graduate School of Linguistics.