Journal of Pragmatics 211 (2023) 47–80

Contents lists available at ScienceDirect

Journal of Pragmatics

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

Processing cost effects of atypicality inferences in a dual-task setup

Margarita Ryzhova*, Vera Demberg

Department of Language Science and Technology, Saarland University, Saarbrücken, 66123, Germany

ARTICLE INFO

Article history:
Received 15 November 2021
Received in revised form 3 April 2023
Accepted 10 April 2023
Available online 4 May 2023

Keywords:
Particularized pragmatic inferences
Informational redundancy
Processing cost
Dual-tasking
Relevance theory
Bayesian analysis

ABSTRACT

Whether pragmatic inferences are cognitively more effortful than processing literal language has been a longstanding question in pragmatics. So far, experimental studies have exclusively tested generalized (scalar) implicatures. Current theories would predict that particularized implicatures should be cognitively effortful – however, this prediction has to date not been tested empirically. The present article contributes to the debate by investigating a specific type of particularized implicature, atypicality inferences, in a dual-task paradigm. In three experiments, we used either a non-linguistic (Experiment 1) or a linguistic (Experiments 2 and 3) secondary task, to modulate the amount of available cognitive resources. Our results show that the strength of pragmatic inferences is largely unaffected by the secondary task, which contrasts with prior predictions. We discuss the implications for traditional and modern accounts of pragmatic processing.

1. Introduction

Language understanding involves recovering the speaker's intended meaning, which might go far beyond the literal semantic meaning of discourse. In pragmatic inferences, listeners often must access information from the situational context, world knowledge or speaker personality and integrate it with the speech signal, see e.g. Degen and Tanenhaus (2015) for an overview. To what extent the integration of the linguistic signal with situational knowledge and world knowledge and subsequent derivation of pragmatic implicatures is cognitively effortful has been a long-standing debate in pragmatics, which has received extensive attention (Noveck, 2001; Bott and Noveck, 2004; Dieussaert et al., 2011; Grodner et al., 2010; Bott et al., 2012; Marty et al., 2013; Cho, 2020). However, the existing research is predominantly concerned with generalized pragmatic implicatures – for example, scalar implicatures, and very little research effort has been devoted to particularized pragmatic implicatures.

A likely reason for this imbalance in experimental testing is that major pragmatic processing theories like the Default account of Levinson (2000), the relevance-theoretic account of Wilson and Sperber (2002) and the constraint-based account of Degen and Tanenhaus (2019) differ in whether they predict generalized implicatures to be effortful or not, but all agree that particularized implicatures should be cognitively effortful.

We do however think that it is worthwhile and necessary to empirically test whether the predictions regarding the costliness of particularized implicatures actually hold.

* Corresponding author. Department of Language Science and Technology, Saarland University, Campus Building C7 2 Room 2.10, 66123 Saarbrücken, Germany.

E-mail address: mryzhova@coli.uni-saarland.de (M. Ryzhova).

https://doi.org/10.1016/j.pragma.2023.04.005

The pragmatic inferences under investigation in this article are not related to lexical scales, but strongly rely on world knowledge about common event sequences (*scripts*; Schank and Abelson, 1975). Examples for a script would be "going to a restaurant", "making coffee" or "going swimming". Comprehenders familiar with these scripts can easily infer the sequence of events that usually happen in these situations.

So-called "atypicality inferences" (Kravtchenko and Demberg, 2015, 2022) arise in conversational contexts as a response to facing utterances that are informationally redundant (or "overinformative") with respect to this commonsense world knowledge.[1] An example is shown in (1) below.

(1) *Today, Lisa went to the swimming pool [...]* ***She brought her swimsuit!***

The sentence "*She brought her swimsuit!*" is overinformative with respect to commonsense script knowledge about going swimming, in that bringing a swimsuit is an integral part of going swimming at the pool.

Following the cooperative principle of conversation (Grice, 1975), encountering a redundant utterance is somewhat infelicitous: cooperative speakers normally tend to be informative, i.e., rather mention unusual events and omit events that are easily inferable (Regneri et al., 2010; Bower et al., 1979).[2]

For a listener, encountering an informationally redundant utterance can trigger a pragmatic inference, which aims to increase their informational utility of the utterance. Kravtchenko and Demberg (2015) showed experimentally that such overinformative utterances as in (1) give rise to the pragmatic inference that the mentioned activity may not normally be performed by the agent in the story. In our concrete example, the inference would be that Lisa does not usually bring a swimsuit to the swimming pool. We believe that the atypicality inference involves two steps: first, the informational redundancy must be noticed, and second, the utterance must be accommodated with the surrounding context by inferring why the speaker might think that the utterance is indeed informative.

All of our stimuli follow the pattern outlined above, in that they are based on a script. They first allude to an everyday activity and then overtly mention an event that can be easily inferred to have taken place (e.g., paying the cashier, when going shopping; eating, when going to a restaurant; or buying a ticket, when taking the subway).

Compared to scalar implicatures, the inferences arising from utterance (1) are highly context-sensitive in that they have been shown to disappear in atypical contexts when, for example, Lisa is presented as a very absent-minded person, or when the mentioned activity is not redundant with respect to world knowledge: an utterance "She brought her children." does not give rise to an inference that the person would not usually bring their children at the swimming pool.

The question we aim to investigate here is whether the pragmatic enrichment being triggered by utterance (1) places significant cost on the comprehender's cognitive system – high enough for these inferences to disappear or be measurably reduced in situations of limited cognitive capacity, e.g., during driving when attention is taxed by another task, or by concurrent tasks which specifically reduce the available working memory capacity.

According to accounts distinguishing between generalized and particularized implicatures, such as Default account of Levinson (2000), inferences like in (1) should be cognitively costly, as the overinformativity needs to be recognized, and alternatives that are not linguistically pre-determined need to be made available (in the above example, one such alternative is that Lisa does not usually bring her swimsuit and maybe even that she is an absent-minded person). Similarly, contextualized accounts of pragmatic processing (such as Relevance theoretic account of Wilson and Sperber, 2002 and Constrained-based account of Degen and Tanenhaus, 2019) would also predict increased processing load, with the following explanation: considering that the atypicality inferences are "contextually supported" only in the sense that commonsense knowledge is necessary to detect the overinformative utterance, there is no direct contextual support for the target atypicality inference, which should carry the processing cost; see Section 2.1 for more detailed explanations.

In the present paper, we methodologically follow previous research in the area of pragmatic processing cost (De Neys and Schaeken, 2007; Fairchild and Papafragou, 2021; Marty et al., 2013; Cho, 2020) by using a dual-task paradigm in our studies. In three experiments, we investigate the role of two different secondary tasks. In Experiment 1, the stimuli are presented auditorily and we use a visuo-motor tracking task that primarily taxes attentional resources and has been linked to some working memory constructs. In Experiments 2 and 3, subjects had to read the materials and we used a reading span task, that more directly taps verbal working memory resources. Thus, the use of two different secondary tasks allows us to investigate the role of two different resources that might be involved in pragmatic processing (attention and verbal working memory, respectively). Contrary to our expectations, we do not find any evidence for processing costs associated with atypicality inferences in either of two settings. This leads us to the conclusion that neither attentional resources nor verbal working memory are substantially taxed by the atypicality inferences we tested here.

The paper is organized as follows. In Section 2.1 we describe in detail predictions of different accounts about the cost of particularized pragmatic inferences compared to generalized ones. Section 2.2 is dedicated to the dual-task experimental methodology previously used in the field for studying pragmatic processing cost. Section 3, presents the predictions of our

[1] Related types of pragmatic inferences have also been shown to arise from other types of redundancy, e.g., when restating facts that are already presupposed (Horn, 1991, 1993, 2014).

[2] We however also note that in the literature on discourse, speaker's overinformativity is not viewed as infrequent (Walker, 1993; Baker et al., 2008). From a position of a speaker, such utterances have been considered either to reflect limitations in speaker's cognitive resources, to be a response to listener's non-comprehension, or simply to carry a narrative function (Walker, 1993; Baker et al., 2008).

study and the trade-off hypothesis of dual-task setup. Sections 4, 5, and 6 describe the experimental results. In Section 7, we conclude the paper and discuss our findings.

The present study hence attempts to fill an important gap in addressing the question whether pragmatic processing is cognitively more effortful than literal interpretation.

2. Background

2.1. Processing cost of pragmatic inferences

Under Grice's model of rational communication, the listener, in their search for the speaker's meaning, is guided by the Principle of Cooperation grounded in four maxims of conversation. In his work, Grice described the procedure the listener should go through to grasp the speaker's meaning (Grice, 1975). However, Grice himself did not consider the proposed steps as a psychologically grounded theory of implicature derivation but rather a philosophical discussion of the steps involved (Zufferey et al., 2019; Mazzarella, 2014).

Later theories were based on Grice's work and attempted to provide more cognitively plausible accounts of communication. In the following section, we discuss the predictions of the Default account of Levinson (2000) and more recent alternative contextualized accounts of Wilson and Sperber (2002) (the Relevance theoretic account) and Degen and Tanenhaus (2019) (the Constrained-based account). We show that, while for scalar implicatures, these two groups (the Default account and the alternative contextualized accounts) differ in their predictions, they agree on the cost for the atypicality inferences.

The proponents of the Default account (Horn, 1972; Chierchia et al., 2004; Levinson, 2000) emphasized the distinction between generalized and particularized conversational implicatures (GCIs and PCIs, respectively – see Huang, 2012; Recanati, 2004, for an overview). GCIs were claimed to be widely independent of context, possessing a clearly defined set of lexical alternatives – as in the "*Mary ate some candies*" example, where the alternatives are tied to {some; all} scale. According to Levinson (2000), these inferences are assumed to occur frequently and are therefore precompiled and virtually free of cost.

In other words, based on this account, people have more exposure to the scalars and their lexical scales, which at the very least reduces the derivational cost of GCIs compared to PCIs (since the set of alternatives in GCIs is available by default) and at the very most makes it automatic and completely cost-free (Chierchia et al., 2004). In turn, PCIs were featured as contextually dependent implicatures, where the set of alternatives is strongly tied to the conversational context. They were viewed as extremely costly, as they are different in each context, and hence the derivation cannot be trained or automatized.

According to PCI/GCI classification, atypicality inferences are clearly particularized, given their high contextual specificity. The alternatives strongly depend on the concrete situations the actor is placed in: either it is Lisa who was uttered to bring her swimsuit when going swimming (and thus the alternative would be that she does not usually bring her swimsuit and maybe she is an absent-minded person) or Lisa who was uttered to pay the cashier when going grocery shopping (and thus she does not usually pay the cashier and maybe she steals food or uses self-checkout automates). The Default-account thus predicts that such inferences should be costly.

In contrast to the Default account, Relevance theory (Wilson and Sperber, 2002) abandons the PCI/GCI distinction and reconciles them in the following way: Under this framework, Gricean maxims are reduced to one single principle of communication relevance. Every utterance raises the expectation of relevance which listeners seek to satisfy by picking up the best hypothesis about what the speaker could mean by saying this utterance. Further, the listener's search for the best hypothesis is determined in terms of cost and benefits. Benefits reflect arriving at the speaker's true meaning, while the cost is the processing effort the listener meets while searching for the meaning. Thus, Relevance theorists emphasized the role of contextual support in both PCI and GCI. An utterance said in a neutral context is predicted to be more costly, as it requires requiring more effort in pragmatic enrichment. However, if a pragmatic interpretation is primed by the context, it can also be less costly than a literal interpretation. According to this line of reasoning, the findings that the scalar implicature similar to "*Mary ate some candies*" can be effortful (e.g. shown in Dieussaert et al., 2011; Antoniou et al., 2016; Bott and Noveck, 2004) is explained from a position that it is not sufficiently primed in the context to avoid the effortful search for relevance.

From the relevance-theoretic perspective (Wilson and Sperber, 2002), processing of atypicality inferences would be predicted to proceed as follows: during a conversation, listeners integrate their representations of the world with the linguistic signal. Their assumptions about the world have different strengths that can dynamically change during the conversation. Crucially, new assumptions can be formed when processing utterances. Before committing an assumption to memory, a cognitive system (they call it a deductive device) checks whether this assumption is already there (e.g., that people usually bring their swimsuits when going to the pool). If the current assumption's strength is not very high, the repetition does not carry any effects other rather than strengthening. However, when a speaker utters information that is already highly predictable from the listener's assumptions (which is the case in bringing one's swimsuit in the going swimming context), to preserve relevance, the listener applies the additional step of processing the repetition and deriving extra contextual effects (that Lisa does not usually bring her swimsuit). Pragmatic enrichment of informationally redundant utterances should hence be associated with some processing costs.

Recently, Degen and Tanenhaus (2019) proposed a new way of understanding pragmatic inferences called the "Constrained-based account." This approach aligns with Relevance theory but additionally proposes a set of factors that influence how we infer meaning. Their paper focuses on scalar inferences and proposes that the likelihood and speed of making a scalar inference depend on how much support it receives from the surrounding context. If the context supports the pragmatic meaning, the probability of making an inference increases, but if the context doesn't support it, the probability decreases.

In the first place, implicature computation depends on the listener's ability to distinguish between different states of the world that can be potentially reported by the speaker. For example, by saying "*Mary ate some candies*", two prospective states of the world are that Mary ate *all* candies and that Mary ate *some but not all* candies. In conversation, the relevant states of the world can be highlighted to interlocutors via the conversational or situational context, which can give rise to a question under discussion (QUD). An example for a QUD in the context of the utterance "*Mary ate some apples*." could for instance be "Did Mary eat *all* apples?". In this case, an alternative for *some* (namely *all*) is explicitly highlighted and the implicature that Mary ate some but not all apples is more likely to arise. Consider on the other hand the QUD "Did Mary eat *any* apples?" – here, the quantifier *any* makes the implicature irrelevant (shown experimentally, for example, in Yang et al., 2018; Kursat et al., 2020).

Furthermore, the computation depends on the properties of the target utterance and its alternatives (what could have been said instead of what the speaker said). According to Degen and Tanenhaus (2019), the target utterance can be characterized by its production cost and/or informativeness. When the listener has access to the alternative utterances (for example via the QUD), they can reason about the cost that the speaker underwent for the chosen formulation compared to the cost the speaker would have undergone if an alternative would have been uttered. Imagine for instance that the word *all* would be more costly to utter than the word *some*. In this case, the less informative utterance "Mary ate *some* candies" would be less likely to give rise to the implicature (shown for reference game setup in Rohde et al., 2012).

The other group of factors that form the conversational context is how the listener judges the speaker. If the listener knows that the speaker does not have a reliable source of information about how many candies exactly Mary ate, for instance, then the implicature is less likely to arise (see Geurts, 2010, for discussion). Next, assuming that the speaker is properly informed, the implicature computation further depends on whether the speaker ostensively tries to be less informative than they could. As a reminder, the use of less informative terms (e.g., *some* compared to *all*) is not considered a lie but only restricts the interpretation to be greater than zero. For example, in the context of Mary being on a diet, the speaker's intention behind saying "*Mary ate some candies*." can be to save Mary's face, i.e., to stay polite with respect to Mary who will be ashamed if other people know how hard the diet is for her, rather than to informatively report that Mary ate all candies. If the listener judges the speaker in this way, then the implicature is less likely to arise. Finally, even when the speaker is maximally knowledgeable and has no intent to be underinformative, the use of a less informative term *some* can be attributed rather to speaker's inability to be properly informative, than to the actual intent to convey an implicature that Mary ate some but not all candies (for example, if the speaker is not a native user of the language, see Fairchild et al., 2020).

Additionally, world knowledge in the form of listeners' prior beliefs can modulate implicature computation. Teresa Guasti et al. (2005) discuss that the experimental materials, where the domain of target world knowledge is not clearly defined, can mask the implicature. For example, in the statement "*Some giraffes have long necks*." the implicature computation depends on a subset of giraffes one has in mind. If the comprehender considers a population of giraffes compared to animals with shorter necks (e.g., lions), then the statement with "some" is underinformative and should be rejected because in fact all giraffes have long necks. However, if the comprehender divides the population of giraffes into baby and adult ones, the implicature in the statement "*Some giraffes have long necks*." is masked: the statement if fairly informative since adult giraffes have longer necks compared to baby giraffes. Another example of the interplay of world knowledge and scalar implicature computation was presented by Degen et al. (2015). They show that scalar implicatures can cause revision of prior knowledge about the world. In their study, subjects were asked to judge how many of the objects (e.g., marbles or feathers) sank in water after hearing an utterance with 'some': "*Some objects sank in water*.". Given the fact that marbles are expected to sink in water more than feathers, one expects prior knowledge about marbles to dominate the implicature (meaning that subjects would answer that all marbles sank). However, as Degen et al. (2015) show, subjects rather tended to make the implicature about marbles (that some but not all marbles sank) – to the same extent as for feathers (e.g., by assuming special properties of the water or the material marbles were made of).

Finally, the implicature computations depends on what information the speaker and the listener share in common ground: does the speaker know that the listener knows their communicative intentions, implied QUD, do they share the information about how they judge each other, do they have any communication conventions, what is known about Mary or, for example, candies etc.

To summarise, when presented out of context, the scalar implicature triggered by "Mary ate some candies." does not receive any contextual support, according to Degen and Tanenhaus (2019): both prospective meanings, *all candies* and *some candies*, are equally likely – neither the question under discussion, speaker's personality, potentially relevant world knowledge, nor common ground are made explicit. Thus, such implicatures should be effortful.

Let us next consider what role, if any, the factors from Degen and Tanenhaus (2019) play in atypicality inferences. In our experimental materials, the conversational context introduces all characters as knowing each other well. Other than that, the context describes an everyday activity and is written in a neutral manner such that none of the actors placed in this context (e.g., the person who speaks about Mary, the recipient of the utterance, or Mary herself) carry any specific properties in terms of their personality, cooperativity or reliability. Similarly, the context does not address the question under discussion – no contextual information is given that can reveal what the speaker wants to convey by uttering the informationally redundant utterance (see Section 4.1 for more detail). Consequently, the set of possible alternatives is not explicit, contrary to scalar implicatures – in our materials, it is completely up to comprehenders' inferencing process. The atypicality inferences are, in fact, triggered by the redundancy of the utterance, but this is different from contextual support of the actual inference:

claiming that the context supports the inference would amount to saying that the mention of going swimming supports the inference that the swimmer in the story does not usually bring their swimsuit, which seems like a crazy claim to make.

Moreover, one might expect even stronger effects of cognitive burden on pragmatic processing here than for scalar implicatures. It is crucial to keep in mind that, compared to atypicality inferences, scalar implicatures, nevertheless, are based on the lexical scales – independently of how the conversational context is defined. In neutral contexts, the set of alternatives is exhaustively formed based on lexical scales. However, in non-neutral contexts, the implicature is derived still based on the scale – albeit accessed in context (Foppolo et al., 2021). In turn, the note raised by Degen and Tanenhaus (2019) about the contextual factors is especially crucial for the particularized pragmatic inferences which arise very situatively, in unique contexts. The set of alternatives consequently might be hard to properly constrain in the listener's mind which can lead to a processing delay or reduced rate of the inferences. This note is one more reinforcement for examining implicatures that do not underlie lexical scales.

On the other hand, one should not expect that contextual support is a cure-all factor for successful pragmatic processing. Following the propositions of Relevance theory, anything that can be relevant for inferencing (according to Degen and Tanenhaus (2019), either coming from the properties of alternatives, speaker's personality or anything else) first needs to be retrieved and then assessed (Wilson and Sperber, 2002). Thus, the inference-relevant information might be ignored or go unnoticed when comprehenders are placed in a situation where they do not have enough cognitive resources.

In the present set of studies, we use a dual-task paradigm to manipulate the amount of available cognitive resources. In Section 2.2, we review how this methodology has been used in the studies of scalar implicatures to test the predictions of theoretical accounts of pragmatic processing.

2.2. Dual-task paradigm in processing cost studies

In the literature on memory, dual-tasking has been shown to effectively tap the executive and attentional resources. The deficit in corresponding sub-components of memory negatively affected performance of the tasks (Holding, 1989; Baddeley and Eysenck, 2010; Baddeley et al., 1991).

The idea is that the secondary task will compete for cognitive resources with the pragmatic processing task, for instance in terms of working memory. If pragmatic processing is effortful in the sense that working memory resources are needed in order to retrieve relevant alternatives and combine the textual information with situational information and world knowledge, pragmatic inferences can be expected to be attenuated when fewer such resources are available during processing.

The idea stems from the Relevance theory claim (and was further broadened in the Constrained-based account of Degen and Tanenhaus (2019) by providing concrete factors forming the contextual support) that when any implicature is not primed in neutral contexts, pragmatic enrichment requires cost for analytical thinking in which working memory is involved (Wilson and Sperber, 2002). In addition, in the studies of individual differences in pragmatic processing, working memory capacity has been found to modulate the success of GCI derivation (Dieussaert et al., 2011; Feeney et al., 2004). Thus, in the absence of sufficient working memory capacity, the rate or strength of inferences should become lower than in the condition with no cognitive burden. In the Default account of Levinson (2000) the same line of argumentation is applicable for PCIs, whereas GCIs are claimed automatic – so, no cost is predicted for them in the dual-tasking setup.

One of the pioneering studies devoted to scalar implicatures by De Neys and Schaeken (2007) examined the rate of pragmatic interpretations in the absence of context. Subjects were presented with underinformative sentences (“*Some tuna are fish*”) and were asked to decide if the sentence was true or false. False responses signified computation of a scalar implicature. As the secondary task, they used a classic spatial storage task, namely the dot memory task. Before encountering an underinformative sentence, subjects were asked to memorize a dot pattern presented in a 3 × 3 grid. After answering the target question, they had to reproduce the pattern. In the low load condition, the dots were vertically or horizontally aligned, which made them easier to memorize. In the high load condition, the matrix contained a rather complex non-linear pattern. The results suggested that under a high load condition, the rate of pragmatic responses was lower than under a low load condition, supporting the claim about the implicatures' costliness in neutral contexts. This finding constitutes direct evidence against the Default account.

In contrast, Fairchild and Papafragou (2021) did not find any effect of load on scalar implicatures, using the same memory dot secondary task. Their experiment contrasted effects of working memory with effects of theory of mind. They found that both working memory capacity and theory of mind correlated with judgments of scalar implicatures but only ToM had a unique effect, while the effect of WMC disappeared once ToM was included as a predictor. Fairchild and Papafragou (2021) suggest that working memory modulates the implicature computation via ToM but not directly – one would need sufficient working memory capacity to carry out ToM computations, or would need sufficient working memory resources to hold results in memory. Similarly, Marty et al. (2013) showed that tapping participant's memory resources interferes with the derivation of scalar implicatures, which speaks against the Default account of Levinson (2000). Participants in their study 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) and reproduce it afterwards. The study did however not find any cost associated to inferences about numerals. In a nutshell, the results of Marty et al. (2013) for scalars support contextualised accounts of pragmatic processing

(Degen and Tanenhaus, 2019; Wilson and Sperber, 2002), while for numerals – the Default account is supported (Levinson, 2000). This shows that even the implicatures that have been, for a long time, claimed conceptually similar – both scalar and numeral implicatures are based on lexical scales and they were classified as GCIs – can still behave differently.

To achieve a stronger interference between the main and secondary tasks, Cho (2020) used a linguistic secondary task (reading span task) in the study of processing cost during online comprehension of scalar implicatures. Their finding also supports the contextualized accounts of pragmatic processing. Based on the analysis of self-paced reading times measured at the final sentence region of underinformative sentences (e.g., *Some birds have wings* and ***beaks***.), there was a difference in reading times between a pragmatically felicitous and a pragmatically infelicitous condition in the no-load condition. This effect however disappeared under increased memory load, indicating that participants may be less likely to recognize the pragmatic problem when under load which in line with contextualized accounts (Degen and Tanenhaus, 2019; Wilson and Sperber, 2002).

To summarise, the results of the previous studies to a large extent disfavor the Default account of Levinson (2000) – under cognitive resources deficit, the implicatures were less likely to arise (as by the rate of pragmatic responses) or took significantly more time for processing (as measured in reaction times). Importantly for us, these results show that the dual-task paradigm is in principle well-suited for detecting cognitive load associated with pragmatic implicatures and also that many types of pragmatic implicatures remain under-researched, with most studies focusing on scalar implicatures. In particular, there are no previous dual-task studies investigating particularized implicatures.

3. Design of the present study and predictions

According to the accounts of pragmatic processing discussed in Section 2.1, the computation of atypicality inferences should be affected by the increased cognitive burden.

To test the predictions, we follow previous studies on scalar implicatures, where the dual-task methodology has been shown to work robustly in detecting the processing cost. In studies of scalar implicatures, a dual-task design has repeatedly shown an interaction between pragmatic processing and load placed on working memory. In the present study, we test whether the same applies to particularised pragmatic inferences triggered by informational redundancy in utterances. Our hypothesis here refers to a trade-off in performance between the secondary task and the main task of interest.

Following the research on multitasking (see, for example, the studies of Li et al., 2001; Fairs et al., 2018; Plummer and Eskes, 2015, in different domains), one can expect that, in a dual-tasking setup, any of the two tasks can be prioritized for processing. Thus, we hypothesize that since in our study, cognitive limitations do not allow to preserve the same level of performance in both tasks, either of the two can be sacrificed for the sake of the other.

In Experiment 1, we used a dot tracking task as a secondary task. The dot tracking task is not a memory-specific task, but is a visuomanual pursuit tracking task that requires continuous attention to a moving target which needs to be followed using the mouse cursor. It thus draws on central attentional resources and has in the past been used simultaneously with language comprehension tasks (e.g., Demberg and Sayeed, 2016). The distance between target and mouse cursor can additionally be used to assess the level of attention expended to the task and detect distraction – see Section 4.2, for more details. To achieve stronger interference between the tasks, in Experiments 2 and 3, we use a linguistic working memory task, the reading span task (Scholman et al., 2020), similar to the one used in Cho (2020).

Thus, in the case of limited processing resources, we predict that the effect size of pragmatic inferences should be substantially reduced in the dual task condition, compared to the single task condition. Considering that in Experiments 2 and 3, the secondary task also requires language processing, we expect even stronger cognitive load effects than in Experiment 1 due to interference of the main and secondary tasks. It is of course also possible that subjects will prioritize the language comprehension task over the secondary tracking or memory task. This should then be measurable in terms of a reduced performance on the secondary task, especially in the condition where the pragmatic inference is drawn.

4. Experiment 1: Atypicality inferences while performing a visuo-motor tracking task

4.1. Materials

The materials for this experiment were taken from Kravtchenko and Demberg (2015) and read out aloud by a native speaker of American English. We selected the 20 items with highest pragmatic effect (defined as the difference in ratings between story conditions) in that study – see Tables A.12 and A.13 in Appendix A, for a full list of stories. Each story consists of a context and an informationally redundant (IR) utterance. The context defines the topic (e.g., *going swimming*, *grocery shopping*) and introduces story characters (2–3 characters per story), see Table 1 part a. Next, one of the story characters mentions an activity which is highly predictable in the context of scenario e.g., *bringing the swimsuit*, *paying the cashier* (bolded utterance in Table 1 part b). The redundancy of the target utterance was normed previously as part of stimuli construction in Kravtchenko and Demberg (2015). The informationally redundant utterances were recorded with exclamatory intonation.

Table 1
Experiment 1. Example of the "Going swimming" story and related questions.

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!**"
Questions
How often do you think Lisa usually brings her swimsuit, when going swimming?
How often do you think Lisa usually brings her children, when going swimming?
What does Lisa like to do after work?

We also constructed four filler stories about everyday activities which had similar properties in terms of structure and length, but did not contain any informationally redundant utterances. Instead, filler stories contained utterances that were either a question ("Hey, do you know what time it is?", "So, what are you up to?", "Have you heard the news today yet?") or an event-unrelated statement ("You know, I'm really tired.").

Each story was followed by three story-related questions. The target question aimed to assess participants' judgments about the target activity typicality mentioned in the informationally redundant utterance (*How often do you think Lisa usually brings her swimsuit, when going swimming?*). The control question addressed an activity that is generally non-predictable from the story (*How often do you think Lisa usually brings her children, when going swimming?*). The third comprehension question was about the content of the story (*What does Lisa like to do after work?*). We used the answers to this question for checking whether participants paid attention to the story content.

4.2. Secondary task: Visuo-motor tracking

To manipulate the participants' available cognitive resources, we used a dot tracking task where subjects were instructed to keep the dot inside of a box, while the dot continuously and randomly moved on a computer screen. Subjects could control the box via their computer mouse – see Fig. 1 for an example of participants' screen while performing a tracking task.

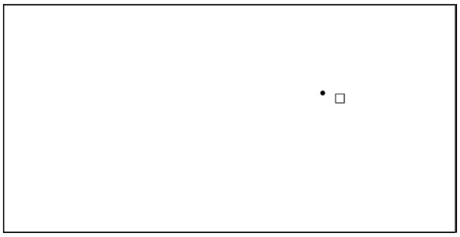

Fig. 1. Experiment 1. Example of participants' screen while performing a tracking task.

The dot tracking task is a type of a visuo-motor tracking task. In the literature on working memory, visuo-motor tracking tasks have been shown to trigger the constructs of working memory, thus affecting the amount of available cognitive and attentional resources (see Baddeley and Eysenck, 2010, Chapter 3, for an overview; Pype et al., 2010).

Our dot tracking task is also very similar in its essence to a continuous tracking and reaction task (the "ConTRe" task, see Mahr et al., 2012). The ConTRe task simulates the driving environment where subjects have to steer the wheel to keep a constantly moving yellow bar in-between of two blue bars. This task was repeatedly and successfully used as a secondary task to elevate cognitive processing load in studies of different linguistic complexity phenomena (see Vogels et al., 2020, for referential processing; Vogels et al., 2018, for semantic surprisal; Engonopulos et al., 2013 for relative clauses).

The dot tracking task and the ConTRe task are equivalent in that they both require constant attention by the participant and provide a continuous measure outcome. However its important advantage over the ConTRe task is that the dot tracking task can be easily used in remote testing. In the current study, we expect that the dot tracking task will primarily tax attentional resources, such that in the dual tasking condition, less attention can be devoted to processing the linguistic stimuli. Alternatively, if participants prioritize the linguistic task, we expect to be able to see that they perform less well on the tracking task during the time of computing the pragmatic inference.

The version of the dot tracking task used in this experiment was implemented according to the example of a dual-task from the website of Cognition Laboratory Experiments, designed by John H. Krantz.[3]

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

Following their design, we controlled the dot via three parameters: maximum angle variation (this describes how much the dot can change direction from moment to moment), speed, and size of the dot. The size of the box was set up equally to the size of the dot. Based on preliminary testing, we balanced the parameters such that the task was challenging but not impossible (size of the dot = 30, dot speed = 600, maximum angle variation = 180). The sampling rate for the dot and the box coordinates was set to 20 Hz (which amounted to taking a measurement every 50 ms).

4.3. Procedure

The experiment started with the instructions where both tasks were explained. Participants were told to listen to the stories carefully and consider their answers to the story-related questions.

Each experiment consisted of only eight trials (four critical trials and four fillers), such that each condition (with-IR vs. without-IR by high vs. no load) was only encountered once by the participants, and no story was encountered in both versions (with and without-IR) by the same participant. This experimental design hence made it impossible for participants to form expectations about what kinds of pragmatic inferences are required in the experiment or develop processing strategies.

In half of the trials, participants thus listened to the stories while tracking the dot with their mouse (high load condition). In the other half of the trials, they only needed to listen to the story. In these trials a fixation cross was displayed in the middle of the screen, and participants were asked to look at it during the duration of the trial (no 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 trial and keep the the dot inside of the box-cursor. After the dot started moving, participants performed 5 s of single-task tracking, before the audio began to play. For the analysis, we annotated the onset of the story as well as the onset of the pragmatic utterance (*She brought her swimsuit!*). The time course of a trial in the high load condition is also illustrated in Fig. 2.

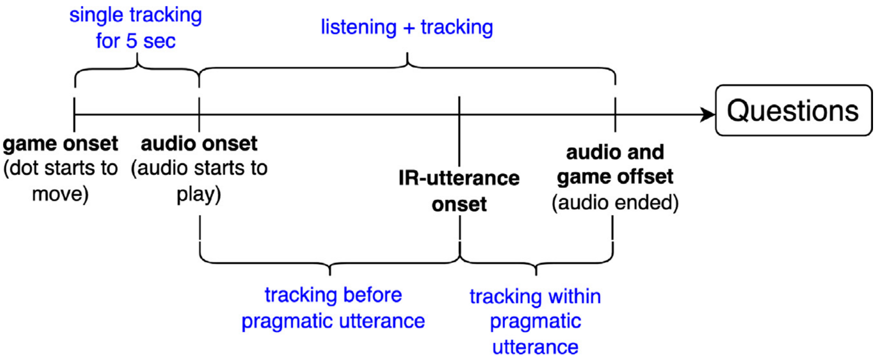
single tracking for 5 sec
listening + tracking
Questions
game onset (dot starts to move)
audio onset (audio starts to play)
IR-utterance onset
audio and game offset (audio ended)
tracking before pragmatic utterance
tracking within pragmatic utterance

Fig. 2. Experiment 1. The timecourse of a trial in the high load condition.

Once the story ended, participants were redirected to a page with target and control questions. To answer these questions, participants had to indicate their estimates using a slider that ranged from 0 ('*Never*') to 100 ('*Always*'). Then, on a separate screen participants were shown a comprehension question which they had to answer in an open form. After all three story-related questions were answered, the next trial began.

The experiment took 15 min on average and participants were paid 2.06 GBP upon completion.

4.4. Data collection

In total, we had 20 different stories. For each story, we had two different versions, one with the informationally redundant utterance, and one without that utterance. We used a 2 (with-IR vs. without-IR) by 2 (high vs. no cognitive load) design. Stories were randomized across 20 experimental lists such that each subject saw each condition only once, with no repetitions of a story topic. Each list thus contained four critical items. We also added four filler trials to each list. For each participant, the order of items was randomized, as well as the order of target and control questions in each trial.

The data collection was conducted to ensure an approximately equal distribution of participants to each list.

4.5. Participants

We recruited 382 eligible participants (mean age = 34 yrs; 60% female) online through 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.

4.6. Analysis

For the analysis of linguistic judgements, we transformed participants' ratings of target activity typicality from their original scale to the open unit interval (0, 1). The transformation consisted of dividing the original ratings by 100 and substituting zero and one values with 0.001 and 0.999 respectively.

The transformed ratings were analysed using a Bayesian mixed effects beta regression model (with a logit link for location parameter μ and an identity link for precision parameter φ), as implemented in the 'brms' package (Bürkner, 2017) in R (R version 4.0.3; 'brms' version 2.15.0). The choice of beta distribution is justified by the nature of the target activity typicality ratings: the ratings are bounded by the experimental design (slider end points), and they exhibit a strong negative skew (histograms of the typicality ratings for each experimental condition are displayed in Appendix B, Figure B.14). As fixed effects, we included story (with-IR vs. without-IR story), load (high vs. no cognitive load), and their interaction. For all parameters, we used the default prior – see Appendix C, Table C.14, for details. The number of iterations was set up to 16,000 with a warm-up of 3000 iterations. Each parameter estimate converged with $Rhat = 1$.

Performance on the dot-tracking task was analysed using by-subject mean tracking deviations as a response variable. We calculated the deviations as the euclidean distance (in pixels) between the dot and the cursor in each timestamp of a trial. For each subject, we calculated the mean tracking deviations per each interval, see Fig. 2. The scaled by-subject mean tracking deviations were analysed using Gamma mixed effects regression models (with an inverse link, as implemented in 'lme4', version 1.1.23 (Bates et al., 2015)). The choice of gamma family was justified by the skewed distribution of the tracking deviations – see Appendix B, Figure B.15. In each model, we included one binary predictor representing two different trial intervals (prior to pragmatic target sentence and pragmatic target sentence) as fixed effects – see Section 4.8 for more details. P-values were obtained using the Satterthwaite approximation for degrees of freedom, as implemented in the 'lmerTest' package, version 3.1.2 (Kuznetsova et al., 2017).

To analyse answers to the comprehension questions, we built a logistic generalized mixed effects regression model of the proportion of correct responses (as implemented in 'lme4', Bates et al. (2015)). Load, story condition and their interaction were used as fixed effects. In all models, all factors were $+0.5/-0.5$ sum coded.

We always started out by fitting models with the maximal random effects structure justified by the design. Thus, for ratings and logistic regression 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). For tracking deviation models, we included by-subject and by-item random intercepts and random slopes for the interval. In the case of non-convergence, we simplified the random effect structure progressively until convergence was achieved (Barr et al., 2013); any model simplifications are stated in the result section.

4.7. Results: pragmatic inferences

First off, we wanted to test whether the pragmatic effect reported in (Kravtchenko and Demberg, 2015) could be replicated: the activity typicality rating for stories including the informationally redundant utterance is expected to be lower than for the stories without informationally redundant utterance. We conducted a Bayesian mixed effects beta regression analysis with by-subject random intercepts and slopes for story and load and by-item random intercepts and slopes for story. Model results are shown in Table 2. We did find evidence of a negative effect of informational redundancy in the story ($CI_{95} = [-0.42, -0.14]$), replicating the effect reported earlier.

Of special interest for our experiment is the interaction between informational redundancy and cognitive load. Here, we would expect the pragmatic effect to be attenuated, i.e., there should be less of a difference in typicality ratings between with-IR and without-IR conditions in the high load condition. However, our experiment did not show any evidence of such an interaction ($CI_{95} = [-0.14, 0.29]$). The histograms of the posterior distributions for the population-level effects are presented in Appendix C (Figure C.18).

Table 2
Experiment 1. Mean estimates, 95% credible intervals, and posterior coefficient probabilities (probability that coefficient > 0), for main effects in the Bayesian mixed effects beta model of linguistic judgements. The mean estimate of the precision parameter φ was equal to 2.53 with $CI_{95} = [2.28, 2.79]$.

Parameter	Mean estimate	95% credible interval	P ($\beta > 0$)
Intercept	1.56	[1.38; 1.75]	1
Story: with-IR	−0.28	[−0.42; −0.14]	0.00025
Load: high	0.02	[−0.09; 0.13]	0.65
Story*Load	0.08	[−0.14; 0.29]	0.76

Fig. 3 displays participants' non-transformed mean activity typicality ratings.

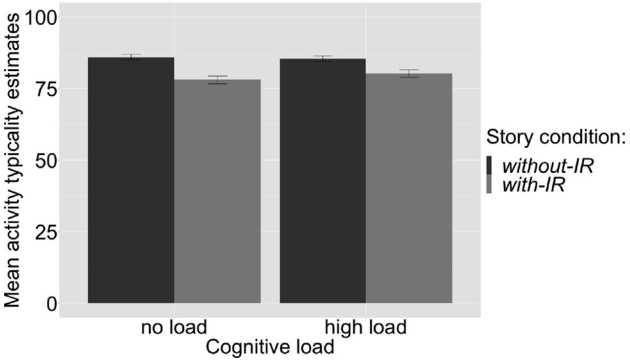

Fig. 3. Experiment 1. Non-transformed mean participants' ratings of the target activity typicality (±SEM) aggregated by story (without-IR vs. with-IR) and cognitive load (no vs. high) conditions.

In addition, we analysed subjects' answers to the comprehension questions. The answers were coded as 1, if the response was correct, and as 0 otherwise. To analyse the proportion of correct responses, we ran a logistic mixed effects regression model. As fixed effects, the model included story, load, and their interaction. The random effects structure consisted of by-subject and by-item intercepts and by-subject slopes for load and by-item slopes for story, see Table 3.

Table 3
Experiment 1. Effect sizes (b), standard errors (SE), z-values, and p-values for the logistic model of the proportion of correct responses to the comprehension questions. Significance codes: *** 0.001 | ** 0.01 | * 0.05.

	b	SE	z	p
Intercept	2.08	0.29	7.28	***
Load: high	−0.32	0.16	−2.04	0.04 *
Story: with IR	0.04	0.23	0.17	ns
Story*Load	0.16	0.3	0.54	ns
Random Effects	*Variance*			
Subject	0.018			
load|Subject	1.36			
Item	0.1			
story|Item	0.3			

The model shows a significant effect of load, suggesting that subjects in the no load condition made fewer mistakes than subjects under high load. This might be considered as an evidence that the dot tracking task indeed did interfere with with language processing.

4.8. Results: tracking deviations

If two tasks, here visuo-motor tracking and language comprehension, compete for attentional resources, effects may be observed on both tasks. If participants prioritize the language comprehension task over tracking, then we would expect (a) reduced tracking performance when comparing single task tracking to tracking during language comprehension and (b) reduced tracking performance when language processing is particularly effortful, i.e., when the pragmatic target utterance is processed compared to non-target utterances.

To address point (a), we compared tracking deviations in single task tracking to tracking deviations during language comprehension.[4] If tracking deviations in dual tasking are higher than tracking deviations in single task mode, this provides evidence that the tasks indeed interfered with one another, and potentially also that participants prioritized the language task over the tracking task. To exclude possible effects of pragmatic processing, the comparison was made on the subset of data in the without-IR story condition. For each interval, we calculated by-subject mean tracking deviations and conducted a Gamma mixed effects analysis. The model showed a significant main effect of task condition ($\beta = 0.04$, $SE = 0.01$, $t = 3$, $p < .003$ ***), showing that people performed less well in tracking while they listened to language at the same time.

[4] To reduce the noise, we excluded the first 2 s of tracking.

Second, to investigate point (b), whether the pragmatic inference was difficult in particular, we conducted an analysis which compared tracking deviations during the non-critical region to tracking deviations in the critical region (tracking during listening before vs. within the pragmatic utterance). The comparison was performed on a subset of data including only the with-IR story condition in the high load condition. The Gamma mixed effects regression model of by-subject mean tracking deviations before and within the pragmatic utterance included by-subject random intercepts[5] and is shown in Table 4. The main effect of interval was significant at $p < .001$, suggesting that participants' tracking deviations were significantly higher in the interval of the pragmatic target utterance than in the interval preceding the onset of the pragmatic utterance (see the aggregated by-subject mean tracking deviations in Fig. 4).

Table 4
Experiment 1. Effect sizes (b), standard errors (SE), t-values, and p-values in the Gamma mixed effects model (with inverse link) of tracking deviations in dual tracking intervals before vs. within pragmatic utterance. Significance codes: *** 0.001 | ** 0.01 | * 0.05.

	b	SE	t	p
Intercept	1.78	0.04	45.43	***
Interval: before	0.05	0.01	3.57	***
Random Effects	*Variance*			
Subject	0.24			

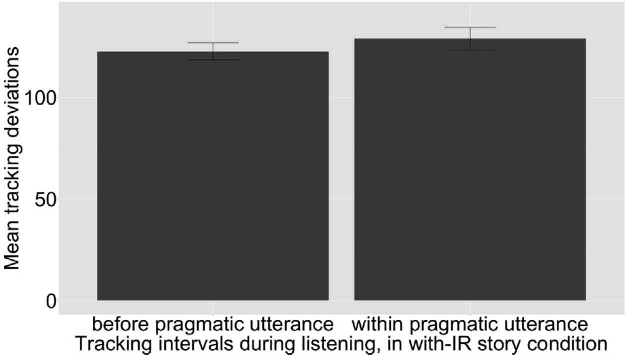

Fig. 4. Experiment 1. Non-transformed by-subject mean tracking deviations (±SEM) aggregated by the interval (before vs. within the pragmatic utterance).

4.9. Discussion

In Experiment 1, we investigated processing cost of particularized pragmatic inferences using a dual-task design where subjects were instructed to listen to the stories with or without an informationally redundant utterance, while performing a visuo-motor tracking task.

We replicated the overall finding reported in Kravtchenko and Demberg (2015) in auditory settings: using Bayesian mixed effects analysis, we found evidence that participants' activity typicality ratings were lower when the predictable activity was mentioned explicitly in the story, in comparison to when it was not. Hence, participants accommodated informational redundancy by lowering their beliefs about the target activity typicality.

A key finding of Experiment 1 is a statistically significant effect of informational redundancy on tracking deviations. In the dual-tasking mode, tracking deviations were significantly higher during the pragmatic target utterance than during literal language processing. This indicates that subjects might have recruited extra cognitive resources in order to process material that elicits a pragmatic inference, directing attention away from the tracking task and hence leading to larger tracking deviations.[6]

In contrast with studies of scalar implicatures (De Neys and Schaeken, 2007; Marty et al., 2013; Cho, 2020), we did however not find a strong effect of cognitive workload on the derivation strength of particularized pragmatic inferences. This might be because participants potentially prioritized the language comprehension task over the tracking task. As there were

[5] Inclusion of by-subjects slopes is not warranted by the design, as there is only one data point available for each subject in each interval.

[6] As was pointed out by an anonymous reviewer, an alternative explanation for the differences between the tracking deviations could be the exclamatory intonation in the region of the pragmatic utterance. To disentangle possible effects of prosody from effects of pragmatic processing on tracking deviations, a future follow-up study could repeat the experiment with a condition where stimuli are recorded in the neutral intonation – see also General Discussion on the role of the utterance's framing in atypicality inferences.

differences in terms of choice of task compared to earlier work which used working memory tasks as a secondary task, we decided to conduct a second experiment using a linguistic WM task, the reading span task, similar to Cho (2020). We predict that the linguistic nature of both tasks will avoid prioritization of the pragmatic task over the secondary task, and allow us to observe reduced levels of pragmatic inferences under load.

A possible limitation of Experiment 1 is also that the sample size might have been insufficient to detect a significant interaction (we did observe a small tendency in the predicted direction, with 75% of posterior samples lying in the predicted > 0 direction). In Experiment 2, we address this by conducting a power analysis based on the observed main effect size of Experiment 1 (Gelman, 2018). Based on the power analysis, we decided to collect approximately 1800 subjects – see details in Appendix D.

5. Experiment 2: Atypicality inferences while performing a reading span task – high vs. low load

5.1. Materials

Materials were taken from Experiment 1. The target and control questions were identical to those used in Experiment 1. For consistency, the comprehension questions in Experiment 2 were rewritten such that the answer could also be given on a continuous scale (see updated stories and comprehension questions in Tables A.12 and A.13, in Appendix A). We also added the fourth filler question about the story characters (*How often do you think Jen and Harvey usually see each other at the pool?*) to each story.

To answer the story-related questions, participants had to put a slider at a position on a scale that best reflected their response. Identically to Experiment 1, the scale ranged from zero (‘Never’) to one hundred (‘Always’).

In order for the stories to contain a clear answer to the new comprehension questions, we modified the context of each story. In Table 5, one can see an updated “Going swimming” story with a new comprehension question “*How often do you think the swimming pool nearby Lisa's office is open, when she finishes working?*”. The number of correct “always” and “never” answers was balanced.

Table 5
Experiment 2. Example of the “Going swimming” story and related questions.

a. Context

Lisa likes to go swimming at a nearby pool after work, as they are always open when her working day is over. 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!**”
Questions
How often do you think Lisa usually brings her swimsuit, when going swimming?
How often do you think Lisa usually brings her children, when going swimming?
How often do you think the swimming pool nearby Lisa's office is open, when she finishes working?
How often do you think Jen and Harvey usually see each other at the pool?

5.2. Secondary task: reading span task

The secondary task was the reading span task. It consists of reading a number of sentences and memorizing their last words for later recall. Sentences in this task are presented in sets of one to four sentences in a row. Each sentence was shown on a separate screen. In the four-sentence-condition, the subjects hence needed to answer four acceptability judgment questions and remember four words. The manipulation of having sets of 1–4 sentences allowed us to manipulate the level of memory load on subjects. To ensure subjects read the whole sentence and not just the last word, they also had to judge whether each sentence was acceptable or not.

The materials for our study were taken from (Scholman et al., 2020) and consisted of one hundred sentences. Sentence length varied from eight to thirteen words. Forty-eight sentences contained a verb that required an animate subject (e.g., escape, forget) and fifty-two sentences contained a verb that required an animate object (e.g., fascinate, impress). Fifty-two sentences were acceptable. Unacceptable sentences were formed by inverting the animacy of the subject and object noun phrases. Below, one can see an example of an acceptable sentence with an animate subject verb (2), acceptable sentence with an animate object verb (3), and unacceptable sentence with an animate subject verb (4).

(2) It was the elephant that escaped from the zoo.
(3) It was the building that impressed the architect.
(4) It was the lawyer that disturbed the phone.

In total in each experimental list, subjects saw 20 different sentences[7] (see Section 5.4 for more details about the lists). For each subject, we calculate the mean last word recall score (hereinafter, the LW-recall score – the number of words recalled correctly overall in the experiment) and the acceptability score (the number of sentences judged correctly overall in the experiment).

The reading span task is a popular verbal working memory measure. In a dual task setting in pragmatic processing, it has previously also been used in another study (Cho, 2020). However, the format of the task was different from our version: it asked the participants to remember the last words of sentences including the critical pragmatic items in the high load condition; the low load condition did not ask them to remember any words. The sentence encoding stage was separated from word recall by a simple arithmetic question. In contrast to our study, participants in Cho (2020) were given feedback for their answers, and there was no time limit for reading/encoding the sentences.

5.3. Procedure

Experiment 2 consisted of reading the stories with or without the informationally redundant utterance and answering story-related questions. The reading span task was used as a secondary task in this experiment to manipulate the amount of available cognitive resources.

Prior to the beginning of the experiment, participants were instructed to avoid using any tools to help them remembering the stories or words. In addition, the sentences were shown on the screen for a limited amount of time which was calculated based on the performance in the training session where subjects were asked to judge only the acceptability of sentences. The mean decision time was used in the main experiment as a threshold after which a 'timeout' message appeared. The timeout decisions were treated as incorrect in the later analysis.

In each experimental trial, participants first read one of the sets of one to four sentences and judged their acceptability. After judging all sentences in a set, participants were shown a story from the primary task, which they had to read carefully. Next, subjects answered the four story-related questions. After participants answered all the questions, they were asked to write down the last words of the sentences they saw before the story. In recall, participants were instructed to follow the original order in which the sentences were presented. The time course of a trial in the high load condition is illustrated in Fig. 5.

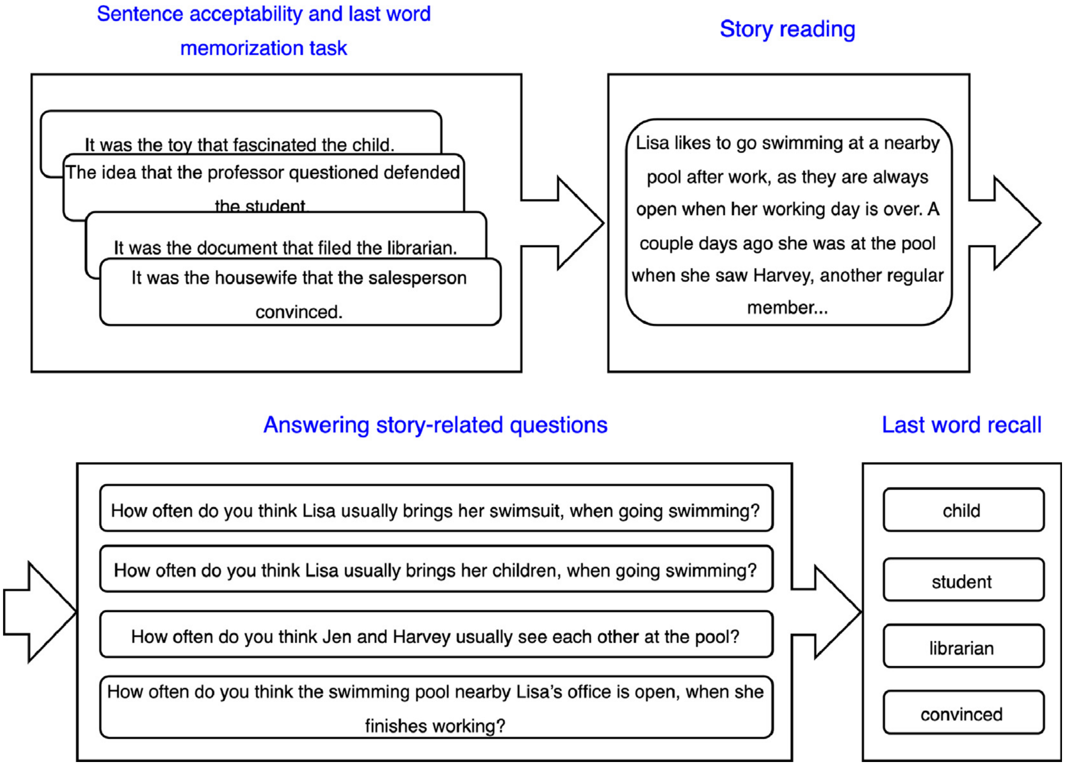
Sentence acceptability and last word memorization task
It was the toy that fascinated the child.
The idea that the professor questioned defended the student
It was the document that filed the librarian.
It was the housewife that the salesperson convinced.
Story reading
Lisa likes to go swimming at a nearby pool after work, as they are always open when her working day is over. A couple days ago she was at the pool when she saw Harvey, another regular member...
Answering story-related questions
How often do you think Lisa usually brings her swimsuit, when going swimming?
How often do you think Lisa usually brings her children, when going swimming?
How often do you think Jen and Harvey usually see each other at the pool?
How often do you think the swimming pool nearby Lisa's office is open, when she finishes working?
Last word recall
child
student
librarian
convinced

Fig. 5. Experiment 2. The timecourse of a trial in the high load condition.

[7] The sentences in each list were equally assigned to each of the above conditions from (Scholman et al., 2020). However, we did not aim to report any differences in the recall rate or the acceptability rate based on these conditions.

5.4. Data collection

We again constructed 20 experimental lists, where each contained one item from each of the four experimental conditions (story type: with-IR vs. without-IR by load conditions: high vs. low load). Experimental lists were designed such that each subject never saw the same item in different conditions, just like in Experiment 1. We also included four filler stories that did not contain an IR-manipulation to each of the lists.

In the low load condition, the set size in the secondary reading span task was one, while in the high-load condition, set size was four. For half of the filler stories, the secondary task set consisted of two sentences, while for another half it consisted of three sentences. Thus, each participant in total saw twenty different sentences from the secondary task and read eight stories in the main linguistic task.

The order of stories and the order of story-related questions was randomized for each subject.

The experiment took 20 min on average and participants were paid 2.73 GBP upon completion.

5.5. Participants

Participation in Experiment 2 was open on the crowdsourcing platform Prolific to those workers who stated English as their native language, who had not taken part in a previous data collection with these items, and who had an approval rating of > 95%.

The data collection happened in several stages in the period of August 2020–August 2021. For the main analysis, the data from 1941 participants were available.

As a part of the data cleaning procedure, we removed from any further analysis 154 subjects who recalled less than 3 words correctly (to avoid analysing misunderstanding of the task or possible technical difficulties). 7 subjects were removed based on their answers to the questionnaire after the study. They had reported English as their non-native language or used other means to help them remembering last words of the sentences.

After data cleaning, we had the data from 1780 subjects. This data forms the basis for the analyses of Experiment 2 (mean age = 34.3 yrs, sd = 12.4; 71% female; mean by-subject acceptability score = 18.2, sd = 1.9; mean by-subject last word recall score = 14.3, sd = 4.4).

5.6. Analysis

For the analysis of linguistic judgements in the target question, we conducted a Bayesian mixed effects beta regression. Histograms of the typicality ratings for each experimental condition are displayed in Appendix B, Figure B.16. The procedure was identical to Experiment 1 as described in Section 4.6 with the following changes. In Experiment 2, we increased the number of iterations to 50,000. This was done to ensure computational stability of the Bayes factor that we calculated for comparing a model with vs. without the story:load interaction. For the same reason, the random structure of the models of linguistic judgements was simplified. We removed by-subject random slopes for story and load. By our experimental design, we had just 2 data points per these parameters which turned out to negatively influence the number of iterations needed to obtain a stable Bayes factor in model comparison.

The answers to the comprehension questions that implied a negative correct response were transformed to a positive scale by subtracting the original rating from the maximum rating of one hundred. Thus, higher ratings in transformed comprehension questions signified higher correctness of the responses (see Figure B.17 in Appendix B). Further, transformed ratings were mapped to the open unit interval $(0, 1)$ as described in Section 4.6. To analyse the transformed ratings in comprehension questions, we built a generalized mixed effects beta regression model as implemented in 'glmmTMB' package (version 1.0.2.1) in R (Brooks et al., 2017). As fixed effects, the model included the load, story, and their interaction. We used maximal random effects models (Barr et al., 2013); in case of non-convergence, we progressively simplified the models as described in Section 4.6.

For the analysis of performance in the secondary task, we built a generalized mixed effects binomial model as specified in 'lme4' package (version 1.1.23) in R. We tested whether the proportion of correctly recalled words in a trial was affected by the presence or absence of the IR-utterance or whether there was a significant effect of story redundancy and load interaction. See details in Section 5.8.

In all models, all factors were $+0.5/-0.5$ sum coded.

5.7. Results: typicality ratings

A Bayesian mixed effects beta regression model with by-subject random intercepts and slopes for story and load and by-item random intercepts and slopes for story was built for participants' ratings of target activity typicality – see Table 6. There was evidence for a negative effect of informational redundancy – subjects showed lower ratings of target activity typicality when the IR-utterance was present in the story compared to stories without informational redundancy (Fig. 6). Most of the posterior samples differed from zero in the predicted direction ($P(\beta < 0) = 0.998$). This replicates the main atypicality inference effect. We also again did not find any evidence for an effect of the interaction ($\beta = 0$, $CI^{95} = [-0.1, 0.1]$).

Table 6
Experiment 2. Mean estimates, 95% credible intervals, and posterior coefficient probabilities (probability that coefficient > 0), for main effects in the Bayesian mixed effects beta model of linguistic judgements. The mean estimate of the precision parameter φ was equal to 2.02 with $CI_{95} = [1.94, 2.11]$.

Parameter	Mean estimate	95% credible interval	P (β > 0)
Intercept	1.36	[1.2; 1.53]	1
Story: with-IR	−0.28	[−0.46; −0.1]	0.002
Load: high	0.02	[−0.04; 0.07]	0.72
Story*Load	0.00	[−0.1; 0.1]	0.5

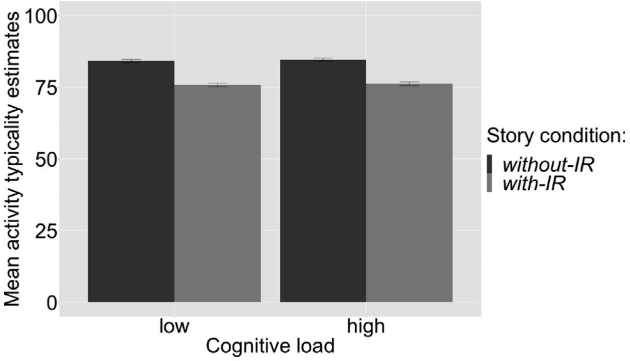

Fig. 6. Experiment 2. Non-transformed mean participants' ratings of the target activity typicality (±SEM) aggregated by story (without-IR vs. with-IR) and cognitive load (low vs. high) conditions.

The histograms of the posterior distributions for the population-level effects are presented in Appendix C (Figure C.19).

The analysis of participants' ratings in the comprehension questions showed significant effects of Story and Load predictors. A generalized mixed effects beta regression model is displayed in Table 7. Subjects' ratings in the high load condition were significantly lower than in the low load condition suggesting that under high load, subjects were less attentive to the stories than under low load ($\beta = -0.07$, $p = .01$). In addition, a marginally significant effect of Story ($\beta = -0.05$, $p = .06$) might suggest that subjects' ratings were less correct when they read stories with IR-utterances compared to stories without the IR utterance – see Fig. 7.

Table 7
Experiment 2. Effect sizes (b), standard errors (SE), z-values, and p-values for the generalized mixed effects beta regression model (with logit link) of the transformed ratings in comprehension questions. The dispersion parameter for beta family was estimated as 1.71. Significance codes: *** 0.001 | ** 0.01 | * 0.05.

	b	SE	z	p
Intercept	1.43	0.05	28.06	***
Story: with IR	−0.05	0.03	−1.85	.
Load: high	−0.07	0.03	−2.56	*
Story*Load	−0.03	0.05	−0.58	ns
Random Effects	*Variance*			
Item	0.05			
Load|Item	0.0009			

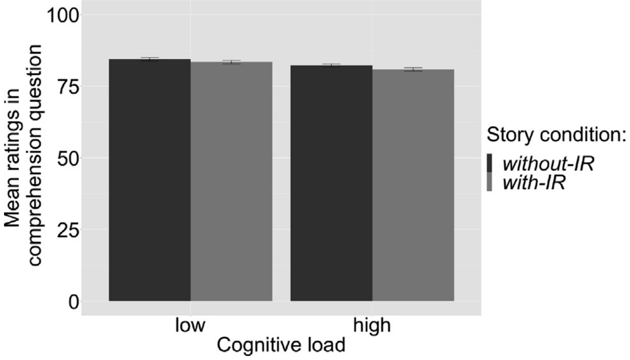

Fig. 7. Experiment 2. Transformed mean participants' ratings in comprehension questions (±SEM) aggregated by story (without-IR vs. with-IR) and cognitive load (low vs. high) conditions.

5.8. Results: recalled words

The mean proportion of correctly recalled words in the low load condition, when one word had to be recalled, was higher (86%) than in the high load condition (66%), when four words needed to be remembered, see Fig. 8 and Table 8 ($b = -0.08$, $p < .05$). We were particularly interested in whether the proportion of correctly recalled words would differ between the story redundancy conditions, as a main effect of story redundancy, or an interaction between load and story. Following the trade-off hypothesis, we analysed the proportion of correctly recalled words. A generalized mixed effects binomial regression model included story and load conditions as well as their interaction as fixed effects. The random structure included by-subject and by-item random intercepts. However, we found no main effect of story redundancy, and the interaction was also not statistically significant. The results are presented in Table 8.

Table 8
Experiment 2. Effect sizes (b), standard errors (SE), z-values, and p-values in the binomial mixed effects model (with logit link) of the proportion of correctly recalled words. Significance codes: *** 0.001 | ** 0.01 | * 0.05.

	b	SE	z	p
Intercept	1.58	0.07	21.51	***
Story: with-IR	−0.08	0.06	−1.48	ns
Load: high	−1.63	0.06	−28.28	***
Story*Load	0.13	0.11	1.23	ns
Random Effects	*Variance*			
Subject	1.92			
Item	0.08			

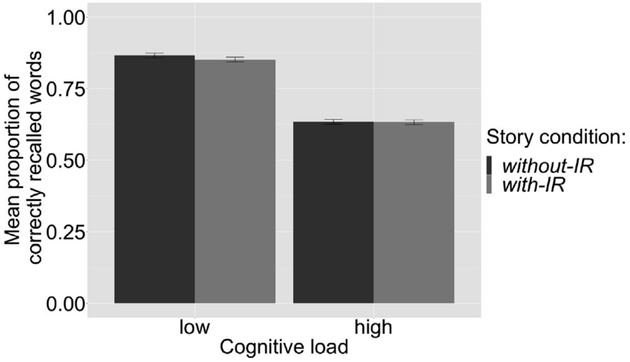

Fig. 8. Experiment 2. Mean proportion of correctly recalled words in a trial (±SEM) aggregated by story (without-IR vs. with-IR) and cognitive load (low vs. high) conditions.

5.9. Bayes factor analysis

In Experiment 2, no evidence towards the processing cost of atypicality inferences was found, as well as no trade-off effects with the secondary task (in terms of the proportion of recalled words). In this section, we will try to quantify the amount of evidence in favour of the null hypothesis, i.e., a model that does not include the story:load interaction to the alternative hypothesis where load is assumed to influence the strength of pragmatic responses, and thus the interaction is present in the model.

Following a Bayesian analysis workflow of Schad et al. (2021), firstly we set up four different priors (default, non-informative, weakly informative, and more informative) to check for the stability of the Bayes factor and whether it critically differs depending on the strength of assumed interaction effect. As a default prior, we chose a flat prior coming from a ‘brm’ function in R. For the other three priors, we specified only the prior for the fixed effects (intercept, story, load, and their interaction) and for the precision parameter φ. A full set of priors used in Experiment 2 is shown in Appendix C, Experiment 2 (Table C.15).

The weakly informative and more informative priors were based on the posterior estimates from Experiment 1 (see Table 3; Figs. 9 and 10 show parameter distributions in the case of more informative and weakly informative priors, respectively). After a visual inspection of posterior distributions, we chose a normal distribution to represent the prior for each of the above parameters (see Appendix C, Experiment 1, Figure C.18). The μs for Intercept and Story type were set to the mean estimates obtained in Experiment 1, as they were the most stable effects we repeatedly obtained in pretest studies. The distributions for load and story:load interaction were centered around zero. The σ estimates for each of the fixed effects in both weakly informative and more informative priors was set close to the corresponding confidence intervals in Experiment 1 but such that some variability still would be allowed (less variability in more informative prior and more variability in weakly informative prior). The parametrization in the non-informative prior for each of the fixed effects was set up to $N(0, 10)$ thus allowing for a wide range of different alternative hypotheses. The parametrization of the prior for φ was chosen to take into account that our data were expected to the skewed to the right to different degrees.

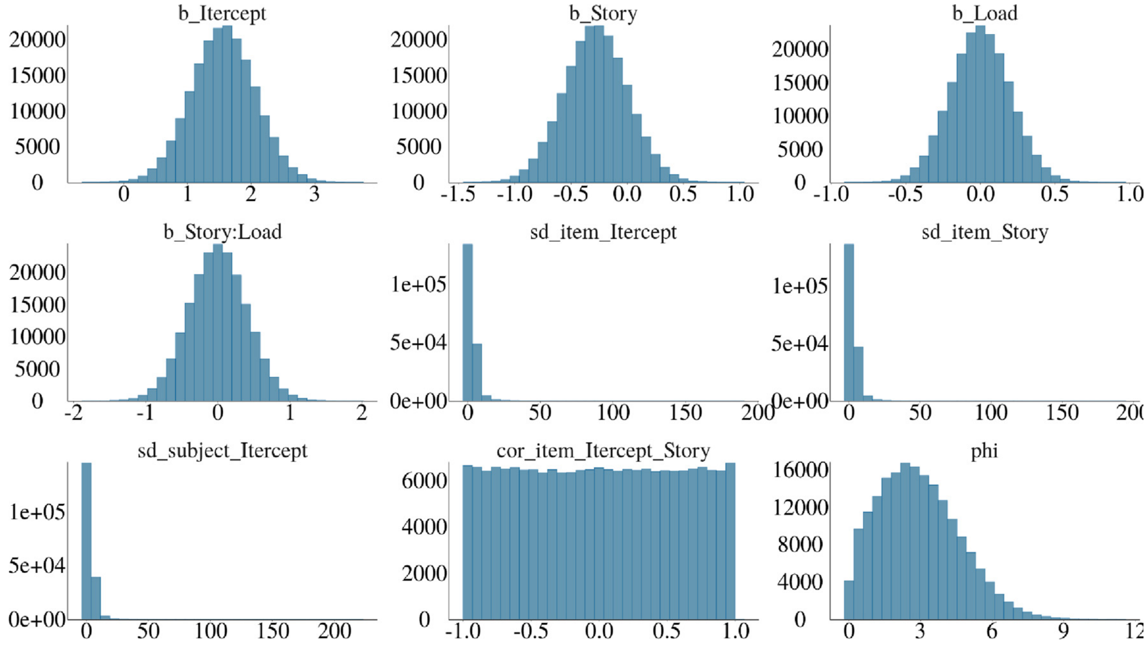

Fig. 9. Experiment 2. Distribution of more informative prior for a set of fixed effect; Prior for other parameters were assigned by default.

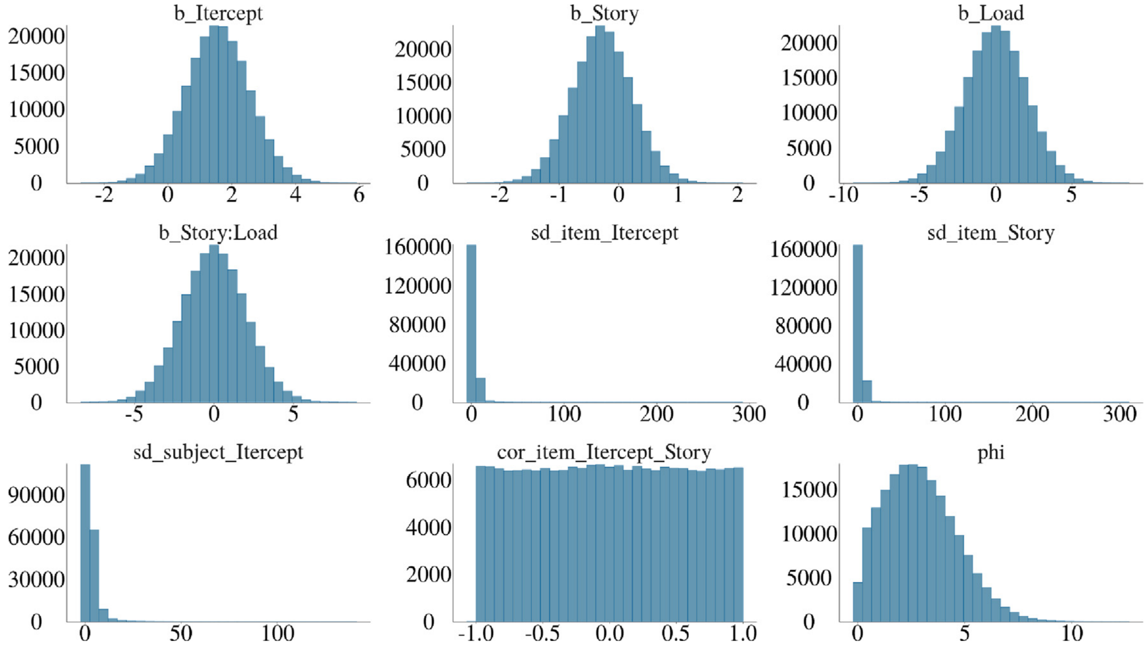

Fig. 10. Experiment 2. Distribution of weakly informative prior for a set of fixed effect; Prior for other parameters were assigned by default.

Note, that posterior estimates did not differ with different priors.

The full model had the same predictors as the model in Table 6. The null model differed from the Bayesian mixed effects beta regression model presented in Experiment 2 only by the absence of the interaction term. For each pair of null and full models we repeated the Bayes factor estimation seven times, to check for stability. The most important observation is that the Bayes factor is always in favour of the null hypothesis, independent of the exact settings of the prior. However, the strength of the estimate in favour of the null hypothesis does depend on the settings of the prior. Our observations are consistent with (Schad et al., 2021) and (Rouder et al., 2009): priors that favour the null hypothesis or smaller effects of the interaction exhibit a smaller odds ratio in favour of the null hypothesis compared to priors that allow for larger interaction effects. Thus, for the weakly informative prior the mean BF01 (evidence in favour of the null over the full model) was equal to $m = 26.87$, $sd = 9$, while for the more informative prior it was $m = 7.25$, $sd = 1.9$. When the prior allowed for a variety of of alternative hypotheses for the interaction (which were not supported by the data) the evidence in favour of the null model increased drastically: $m = 217.4$, $sd = 39.4$.

5.10. Discussion

In Experiment 2, we tested the cost of atypicality inferences by modulating the load via a reading span task. Compared to Experiment 1, using a secondary task of linguistic nature was supposed to increase the interference with the main pragmatic task.

However, there was found no direct evidence for an effect of load on the strength of pragmatic inferences. A Bayesian hierarchical Beta regression analysis of ratings to the target typicality questions included an estimate for a story:load term that was symmetric around zero ($\beta = 0$, $CI_{95} = [-0.1, 0.1]$). As a follow-up, we conducted a Bayes factor analysis where we compared the models with and without the interaction term. The Bayes factor consistently suggested evidence in favour of the null hypothesis. Notably, even the models which in their prior assume already that the effect will be very small favoured the null.

On the other hand, there was found no reduced performance in the secondary task. In general, subjects' mean proportion of recalled words was lower under high load than under low load which justified that conditions of the secondary task varied in difficulty for subjects. However, the number of correctly recalled words was not affected by the presence or absence of the informational redundancy in stories. Thus, under high load, subject did not seem to compensate the same level of performance in pragmatic task by loosing on words' recall.

Finally, consistently with Experiment 1, the load coming from the secondary task influenced ratings to the comprehension questions – subjects' answers were less correct under high load than under low load ($\beta = -0.07$, z-value $= 0.03$, p-value $= 0.01$) but there was no significant effect of story redundancy or story:load interaction.

Considering the linguistic nature of the secondary task in Experiment 2, if the processing of atypicality inferences would be costly, the trade-off between the tasks here should have been more pronounced than in Experiment 1. However, we neither found an effect on pragmatic inferences, nor a compensatory effect in the secondary task performance, in contrast to Experiment 1. One possible explanation for our failure to find an effect in Experiment 2 could be that even our low load manipulation was already sufficiently taxing and distracting, that no difference between the conditions could be found. This is in contrast to Experiment 1, where the easy condition involved no secondary task. We therefore decided to follow-up with a third experiment, which contrasts the high load condition with a no-load condition, where subjects did not have to remember any of the words prior to story reading and question answering.

6. Experiment 3: Atypicality inferences while performing a reading span task – high load vs. no load

6.1. Materials and experimental setup

The experimental materials were taken from Experiment 2. The experimental design and procedure were identical to Experiment 2, with the following changes: Contrary to Experiment 2, the current experiment consisted of two blocks, a single task block and a dual task block. The dual task block contained the high load condition from Experiment 2, i.e., the trials were identical to Experiment 2, and each story was always preceded by four sentences. Thus, each subject saw 16 sentences for which the last word had to be memorized in the study. In the single task block, no preceding sentences were shown, and no memory task was required – subjects were only instructed to read the stories and answer story-related questions. The order of blocks was randomly assigned to each participant as they entered the study. Each block consisted of four stories: one story in with-IR condition, one story in without-IR condition, and two fillers. The order of stories within each block was randomized.

The experiment took 15 min on average and participants were paid 2.58 GBP upon completion.

6.2. Participants

842 subjects completed the study (mean age = 42.2, sd = 13.8; 63.9% female, one subject preferred not to report; mean by-subject acceptability score = 14.4, sd = 1.5; mean by-subject last word recall score = 8.7, sd = 5). The requirements for participation were the same as in Experiment 2: eligible participants stated English as

their native language, did not take part in the previous studies with the same materials and had an approval rating of 95%.

The data cleaning procedure was the same as in Experiment 2.133 subjects, who recalled less than three words in the whole study (out of 16 possible words), were removed from any further analyses. The proportion of removed subjects based on this criterion is higher than in Experiment 2. The analysis of their performance and answers to the questionnaire after the study showed that excluded subjects did not report any technical difficulties; in the recall task, they either did not write any words or they wrote random words from the sentences, but not their last words. One possible explanation for such a high number of low performers might be that in Experiment 3, subjects were always faced with four words per trial, while in Experiment 2 they had also trials with one, two or three words. This might have made it easier to cross the threshold in Experiment 2 compared to Experiment 3.

We decided to nevertheless maintain this criterion for excluding participants, as they might have focused only on the story reading task, and in that case would be less likely to exhibit any effect of load on pragmatic inferences. We note though that a post-hoc analysis of excluded participants showed no difference in their ratings between the no load and the high load condition; adding these participants to the sample would hence not change results (see section 6.3).

After data cleaning, 709 subjects (mean age = 42, sd = 13.8; 64.2% female, one person preferred not to report; mean by-subject acceptability score = 14.5, sd = 1.5; mean by-subject last word recall score = 10.2, sd = 3.9) were kept for the analyses.

6.3. *Results: typicality ratings*

A generalized mixed effects beta regression model with by-subject and by-item random intercepts and by-item random slopes for story condition was built to analyze subjects' ratings of target activity typicality – see Table 9 and Fig. 11. As in the previous experiments, we found a significant effect of informational redundancy ($\beta = -0.24$, $p < .01$). We also found a significant effect of load ($\beta = -0.11$, $p < .01$), meaning that subjects' typicality ratings were on average lower under high load than under no load. An interaction between story and load was not significant.[8]

Table 9
Experiment 3. Effect sizes (b), standard errors (SE), z-values, and p-values for the generalized mixed effects beta regression model (with logit link) of linguistic judgements. The dispersion parameter for beta family was equal to 2.25. Significance codes: *** 0.001 | ** 0.01 | * 0.05.

	b	SE	z	p
Intercept	1.5	0.09	17.5	***
Story: with IR	−0.24	0.08	−3.1	**
Load: high	−0.11	0.04	−2.7	**
Story*Load	0.12	0.08	−1.5	ns
Random Effects	*Variance*			
Item	0.13			
Story|Item	0.09			
Subject	0.06			

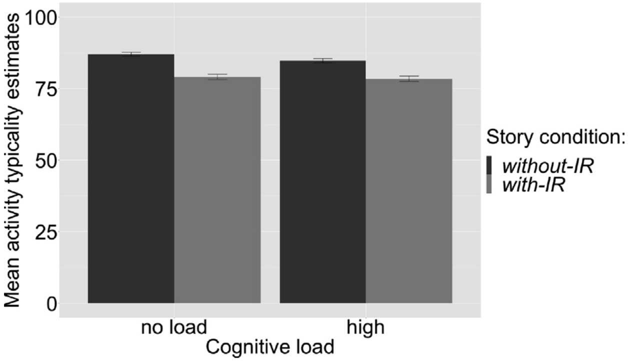

Fig. 11. Experiment 3. Non-transformed mean participants' ratings of the target activity typicality (±SEM) aggregated by story (without-IR vs. with-IR) and cognitive load (no load vs. high) conditions.

[8] Considering that the order of blocks could potentially influence the overall cognitive load of subjects throughout the experiment, we also included the block order and its interactions in the model. The idea behind was that if subjects were firstly exposed to the block with no secondary task, they might have had more resources compared to facing the the same block after the dual task. However, there was no significant 3-way interaction.

The analysis of participants' ratings in the comprehension questions also did not show a significant effect story and load interaction – see Table 10. We found only an effect of load, meaning that subjects were less attentive to a semantic content of the stories under high load, compared to no load – see Fig. 12.

Table 10
Experiment 3. Effect sizes (b), standard errors (SE), z-values, and p-values for the generalized mixed effects beta regression model (with logit link) of the transformed ratings in comprehension questions. The dispersion parameter for beta family was estimated as 1.71. Significance codes: *** 0.001 | ** 0.01 | * 0.05.

	b	SE	z	p
Intercept	1.67	0.06	29.84	***
Story: with IR	−0.07	0.04	−1.62	ns
Load: high	−0.22	0.04	−5.31	***
Story*Load	−0.05	0.08	0.62	ns
Random Effects	*Variance*			
Item	0.05			

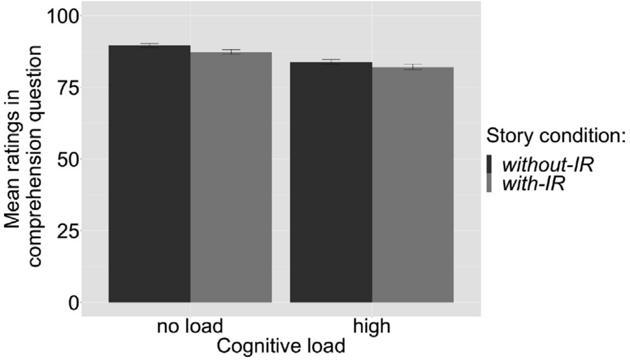

Fig. 12. Experiment 3. Transformed mean participants' ratings in comprehension questions (±SEM) aggregated by story (without-IR vs. with-IR) and cognitive load (no load vs. high) conditions.

6.4. Results: recalled words

The mean proportion of correctly recalled words was equal to 64.6% which was comparable with the recall rate under the high load in Experiment 2 (66%).

A generalized mixed effects binomial regression model of the proportion of correctly recalled words included only story condition. The random structure included by-subject and by-item random intercepts and by-item random slopes for story. There was found no main effect of informational redundancy. The results are presented in Table 11 and Fig. 13.

Table 11
Experiment 3. Effect sizes (b), standard errors (SE), z-values, and p-values in the binomial mixed effects model (with logit link) of the proportion of correctly recalled words. Significance codes: *** 0.001 | ** 0.01 | * 0.05.

	b	SE	z	p
Intercept	0.8	0.07	11.8	***
Story: with-IR	−0.03	0.1	−0.2	ns
Random Effects	*Variance*			
Subject	1.66			
Item	0.02			
Story|Item	0.32			

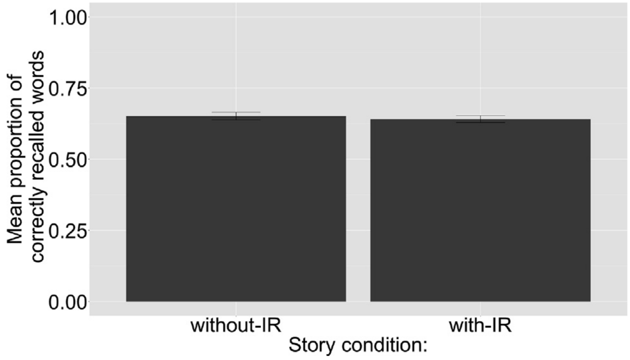

Fig. 13. Experiment 3. Mean proportion of correctly recalled words in a trial (±SEM) aggregated by story condition (without-IR vs. with-IR).

6.5. *Discussion*

Experiment 3 followed up on Experiment 2 by making the manipulation more extreme – instead of comparing low load to high load, we here compared no load to the high load condition. This allows us to test whether the lack of interaction between load and informational redundancy in Experiment 2 might be due to both the high load and low load condition both inducing cognitive load which affects pragmatic inferences, making it impossible to find a difference between these conditions.

However, we did not find a significant interaction between story and load. There was only a significant effect of load, showing that subjects' ratings of the target activity typicality, independently of the presence of an IR-utterance, were on average lower under high load, compared to no load. This means that subjects in general tended to pay less attention to the content of the story when they also had to remember the words. Subjects' ratings in the comprehension questions were also significantly lower under high load than under no load.

We also found no differences related to the order in which the experimental blocks (single task and dual task) were presented. Similarly to Experiment 2, we also did not observe any trade-off effects in subjects' performance in the secondary task: in the high load condition, there was no effect of informational redundancy on the proportion of correctly recalled words. Experiment 3 thus fully confirms the findings of Experiment 2.

7. General Discussion

The experiments reported here were aimed at testing whether a specific type of particularized pragmatic inference, namely an atypicality inference in the context of an informationally redundant utterance, is cognitively costly. Kravtchenko and Demberg (2015, 2022) previously showed that subjects tend to lower their initially high beliefs about the typicality of a highly predictable event (e.g., bringing one's swimsuit at the swimming pool) when this activity is mentioned explicitly. Atypicality inferences are highly context-dependent – in contrast to scalar implicatures, there is no context-independent lexically encoded alternative. Therefore, such inferences should be cognitively costly, according to theoretical accounts of pragmatic processing (Degen and Tanenhaus, 2015; Sedivy, 2007; Wilson and Sperber, 2002).

Following the methods previously used in the studies investigating the costliness of scalar implicatures, we used a dual-task design to manipulate cognitive load. Subjects had to perform two tasks in parallel, thus leaving less resources available for each task compared to when each of the tasks would be addressed separately. One of our predictions was that in a situation of cognitive burden, subjects potentially would not recognize the redundancy and/or not draw any atypicality inferences based on the redundancy. In both cases, the strength of pragmatic inferences would be expected to be lower on average than in the low or no load conditions. In Experiment 1, the load was imposed via a visuo-motor tracking task, where subjects had to continuously track a dot while listening to the stories. In Experiments 2 and 3, the load came from a reading span task. In Experiment 2, subjects had to memorize 1–4 words and recall them after performing the main pragmatic task. In Experiment 3, subjects had to memorize either 4 words (high load) or no words (no load), thus making this experiment more directly comparable to Experiment 1.

None of our experiments showed direct evidence for processing cost associated with atypicality inferences – the strength of the atypicality inferences was equal in the high and low/no load conditions which contradicts the prominent existing accounts of the pragmatic processing. Only in Experiment 1, there was a small effect on the tracking deviation in the pragmatic condition, indicating that subjects might recruit additional resources for processing the pragmatic utterance, and then might compensate the same level of performance in pragmatic processing by sacrificing performance in the secondary task. We note though that it is at the point unclear whether the observed effect is due to pragmatic processing, or due to more low-level processes related to processing the exclamatory prosody.

In Experiment 2, no compensatory trade-off between the two tasks was found. We do not think that the lack of effect can simply be attributed to power issues, as we hired a large number of participants in Experiment 2, enough to detect an interaction effect that is half the size of the main effect of story type with a probability >80%. Our Bayes factor analysis consistently showed evidence in favour of the null hypothesis, i.e., that there is no interaction between cognitive load and pragmatic inference. Considering that in Experiment 2 the low load could have been potentially still strong enough to influence pragmatic processing, we conducted Experiment 3, with no and high load conditions. The results of Experiment 3 were fully consistent to Experiment 2: we did not find any effects of load on pragmatic processing, nor any effects of informational redundancy on the secondary task performance.

A first issue to consider is whether our load manipulations were successful. We believe that this was the case, as we observed a reduction in comprehension answer accuracy in all three experiments when the cognitive load induced by the secondary task was high. Comprehension questions were constructed such that they had clear Yes/No responses based on the text. Participants did not have to, and actually could not, infer the answer to these questions from world or script knowledge.

Secondly, the effect found in Experiment 1 could potentially be an artifact of the exclamatory intonation and in fact might not reflect difficulties in pragmatic processing. In this case, the higher tracking deviations found in the region within pragmatic utterance, compared to the region before, would reflect subjects' surprise to the exclamatory intonation itself. Since in Experiment 2 the stimuli were presented textually, the exclamation mark simply might not have had a comparable influence as the auditory intonational prominence, and this could thus be the reason why no effect on secondary task performance was found in Experiment 2. To rule out this possible explanation, future work could repeat Experiment 1 in a setting where both conditions are using exclamatory prosody, or with an additional condition where the utterances are spoken in neutral intonation. We note however that the effect sizes for atypicality inferences without exclamatory intonation are expected to be lower in that condition, according to previous work by Kravtchenko and Demberg (2022). The experiment would hence require several thousand participants to be well-powered. Finally, an important difference between Experiments 1 and 2 can be the modality in which the stories were presented to the participants. While there are no reported differences in memory recall between listening and reading (see e.g., Rogowsky et al., 2016), it still can differently affect online processing in the situation of cognitive overload.

Next, we will discuss the connection between script knowledge and atypicality inferences in the context of retrieving information from memory. Atypicality inferences are based on script knowledge, which refers to the sequence of events related to everyday activities that are stored in our memory, such as swimming, buying food, and getting hair washed. This type of knowledge is highly crystallized in memory and can be immediately accessed and integrated with context. The literature on world and script knowledge suggests that interlocutors come into a conversation with a set of assumptions about the world, the conversational partner, and other encyclopedic facts of different granularity. These assumptions are progressively updated as the conversation evolves (Hagoort et al., 2004) and new information is introduced (Chwilla and Kolk, 2005). However, listeners cannot hold all assumptions in their working memory, so a search process becomes necessary. Each new piece of information can trigger different sets of assumptions from diverse sources, including perception, short-term memory, and long-term memory. The organization of encyclopedic memory plays a crucial role in the pursuit of relevance, and how information is chunked can either facilitate or hinder the search for accommodating inferences. According to Wilson and Sperber (2002), retrieving relevant information from memory can require more effort than interpreting an utterance, but some pieces of information might be more readily available and easier to retrieve than others.

Thus, one can hypothesize that the overinformativity can be recognized very easily: the script is activated by the story topic and induces a high expectation of the target event; this target event is hence easy to retrieve, and its redundancy can be easily recognized. But could the absence of cost for atypicality inferences be explained this way? A question that requires future research is whether the ratings that people gave were connected to viable interpretations for a belief reduction, i.e., whether and if yes, at what moment, an alternative explanation was formed (i.e., *Lisa doesn't usually bring her swimsuit to the swimming pool* ***because*** *this is a nudist swimming pool*). Such a richer inference of reconciling the atypical event with world knowledge would not seem to be always readily available, and hence should incur a cost. In a recent study that did not include a cognitive load manipulation, we elicited explanation of given ratings from participants. The results, in fact, show that those participants who lowered the typicality rating in the informationally redundant condition often provided such richer explanations (Ryzhova et al., 2022). However, the current studies are not able to tease apart *when* exactly the inferences for accommodation of the informationally redundant utterance happen – immediately when the utterance is encountered, at the end of the sentence, or with a delay at the point in time when the rating is given, or maybe even only when the justification is elicited. We note that in Experiment 1, the question took place in a single task setting, so any difficulty related to the inference which would only take place at question time cannot be measured in the secondary task. In Experiments 2 and 3, on the other hand, the secondary task (memory) was still on-going at question time. In future work, a new experimental design should be used to investigate the relationship between forming alternative explanations and the effects of cognitive load.

We would now like to discuss our findings in the light of two other recent studies, Foppolo et al. (2021) and Fairchild and Papafragou (2021). Foppolo et al. (2021) shows that inference-relevant information, under some conditions, can be obtained while avoiding cost. Using a Picture Selection task, they compared the rate of pragmatic enrichment that children made in classic scalar implicatures (target sentence: *On my birthday cake, some of the candles are burning*; visual scene: a cake with 3/5 burning candles, 0/5, 5/5, and no burning candles) vs. particularized implicatures where to make an inference, children had to judge intentions of the interlocutor based on context (target sentence: *On my bed there is a teddy bear*; visual scene: the bed with a teddy bear and a penguin, only a teddy bear, only a penguin, and an empty bed). The results of this study suggest that contextual cues can be utilized fast enough to avoid the cost of deriving alternatives under some conditions. In their study, contrary to predictions of original accounts, the derivation rate of particularized implicatures was higher than that of scalar implicatures (86% vs. 74% respectively). In addition, the success rate of both implicature types correlated with morphosyntactic competence, while the theory of mind ability correlated only with scalar implicatures but not with particularized implicatures. Foppolo et al. (2021) hypothesize that, under the assumption that children have not yet mastered lexical scales at a level of adults, particularized implicatures, in their study, served contextual cues that helped to grasp the speaker's intentions and made alternatives more visible. While in scalar implicatures, children had to infer the alternative sets based on their mental state reasoning abilities.

We should also consider whether our hypotheses regarding the processes being involved in particularized inferences are correct, or whether a different set of dual tasks, not involving attention or working memory, might have interfered more strongly with drawing atypicality inferences, and thereby would have elicited a dual-task effect. In a recent study, Fairchild and Papafragou (2021) found that individuals with better theory of mind abilities (ToM) were more likely to compute scalar implicatures. More importantly to our results, no evidence of a unique contribution of executive functions (including working memory capacity) was found. We note that in this concurrent study, Fairchild and Papafragou (2021) also failed to find an effect of cognitive load on pragmatic inferences. Fairchild and Papafragou (2021) hypothesize that the executive function might be involved in pragmatic processing only to the extent to which it is recruited in theory of mind reasoning (as such computations might be resources-demanding) and its further consolidation with the rest of contextual information. As pointed out by an anonymous reviewer, there might be a possibility that our results are due to atypicality inferences not requiring theory of mind reasoning – namely, the participant (who acts as an overhearer in the experimental setup) might compute an atypicality inference irrespective of the speaker's or listener's (mutual) knowledge states. In support of this hypothesis, the reviewer considers recent findings in language production that speakers mention atypical events because they cannot be inferred by a generic listener rather than due to the demands of a specific listener. That is why for computing an atypicality inference, the overhearer's mentalizing skills might not be involved in judging the mental states of the story's interlocutors (Brown and Dell, 1987; Grigoroglou and Papafragou, 2019; Lockridge and Brennan, 2002).

However, we think that the hypothesis that atypicality inferences should be completely cost-free and they do not require any ToM should be considered with caution. As we write above, recognising redundancy might, in fact, come at no cost and without involving any mentalizing skills due to the activation of related script. However, we believe that accommodating the informational redundancy and forming rich explanations can not be done without assessing speaker's knowledge state, and this should be addressed in future work. To process explanations about Lisa (e.g., that she is forgetful or that she visits a nudist swimming pool and that's why does not usually bring her swimsuit), the comprehender of the story might need to think about the context and knowledge states of the story characters. E.g., the speaker and the listener mutually know about Lisa's habits that she often forgets to bring her swimsuit to the swimming pool. This fact about Lisa (or any person who goes swimming) is not in world knowledge per se (only the knowledge that people bring their swimsuits is). The overhearer should compute that this information is in the characters' common ground and that mentioning the normally typical course of actions (that she brought her swimsuit) is informative. In fact, our story contexts are carefully set up to make sure that the characters in the stories are not strangers to each other but that they know each other well. Moreover, forming such rich alternative explanations about Lisa might involve other cognitive factors as well, for example, general reasoning ability, as it builds up on complex relations between the characters, traits they might possess, and sometimes society rules (see Ryzhova et al., 2022, for preliminary results).

Finally, we would like to note that, while our results show no effects of load on the pragmatic processing of a particular type of particularized conversational inferences, this conclusion cannot necessarily be generalized to other PCIs. Future studies should investigate the effect of cognitive load in the processing of other particularized implicatures, specifically those that are less dependent on common sense knowledge.

Funding statement

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Funder Id: https://doi.org/10.13039/501100001659, Project-ID 232722074 – SFB1102: "Information Density and Linguistic Encoding".

Declaration of competing interest

The authors declare no conflicts of interest.

Data availability

Data will be made available on request.

Appendix A. Experimental materials

In the following appendix, we provide a full set of experimental materials (stories about everyday situations and corresponding questions — see Tables A.12 and A.13, respectively) used in our experiments.

Stories in the without-IR story condition comprised only of the context, while in the with-IR story condition, it was context + IR-utterance.

Stories in Experiment 1 were identical to those used in the original study of Kravtchenko and Demberg (2015). For Experiments 2 and 3, we changed the stories such that the answers to the comprehension question could also be given on a continuous scale (Table A.12, column “Context in Exp-s 2 and 3”). We highlight in bold the changes we made in the stories with respect to Experiment 1 (Table A.12, column “Context in Exp 1”). The target questions about the target event typicality were identical across all three experiments and the original study of Kravtchenko and Demberg (2015).

Table A.12
Experimental Materials for Exp. 1, 2, and 3.

Story	Context in Exp 1	Context in Exp-s 2 and 3	IR-utterance
grocery	John often goes to the grocery store around the corner from his apartment. Recently, he came home from the store with groceries. When he came in, he saw his roommate Susan in the hallway, and started talking to her about his trip to the store. As he went to the kitchen to put his groceries away, Susan went to the living room, where their roommate Peter was watching TV.	John often goes to the grocery store around the corner from his apartment. Recently, he came home from the store with groceries. When he came in, he saw his roommate Susan in the hallway **preparing for some yoga exercises, as she does all the time at home. John** started talking to her about his trip to the store. As he went to the kitchen to put his groceries away, Susan went to the living room, where their roommate Peter was watching TV.	Susan said to Peter: “John just came back from the grocery store. He paid the cashier!”
restaurant	Mary is a journalist who often goes to restaurants after her interviews. Yesterday, she went to a popular Chinese place. As she was leaving, she ran into her friend David, and they started talking about the restaurant. After they parted, David continued on his way when he suddenly ran into Sally, a mutual friend of him and Mary.	Mary is a journalist who often goes to restaurants after her interviews. Yesterday, she went to a popular **Italian place, where they always have nice live music**. As she was leaving, she ran into her friend David, and they started talking about the restaurant. After they parted, David continued on his way when he suddenly ran into Sally, a mutual friend of him and Mary.	David said to Sally: “I ran into Mary leaving that Chinese place. She ate there!”
feed dog	Jim lives in a shared apartment, where it's his job to feed the dog in the evenings The other day he was feeding the dog some canned food, as his roommate Lucy came into the kitchen, and made herself a snack while chatting with him. Later in the evening, she settled down to watch TV alone with their roommate Carl.	Jim lives in a shared apartment, where it's his job to **always** feed the dog in the evenings. The other day he was feeding the dog some canned food, as his roommate Lucy came into the kitchen, and made herself a snack while chatting with him. Later in the evening, she settled down to watch TV alone with their roommate Carl.	Lucy said to Carl: “Jim was feeding the dog earlier. He threw the can away!”
subway	Jane takes the subway all the time to get around the city. Today she was entering a subway station when she ran into her friend Don, and they took the train together as they were heading in the same direction. Later that day, Don ran into Beth, Jane's sister, on the street.	no changes	Don said to Beth: “I took a train with Jane today. She bought a subway ticket!”
swimming	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.	Lisa likes to go swimming at a nearby pool after work, **as they are always open when her working day is over**. 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.	Harvey said to Jen: “Lisa's here to swim, too. She brought her swimsuit!”
train	Brian takes the train most mornings, although the commute takes a long time. Last week when he was getting on the train, he ran into his old colleague Rachel, and they chatted until Brian got off. When Rachel got to work, she saw Oliver, who also used to work with Brian.	Brian takes the train most mornings, although the commute takes a long time. Last week**, as always, he bought a cup of coffee, and** when he was getting on the train, he ran into his old colleague Rachel. They chatted until Brian got off. When Rachel got to work, she saw Oliver, who also used to work with Brian.	Rachel said to Oliver: “I saw Brian on the train this morning. He got off at his stop!”
work	Laura works as a software engineer at a large company. A couple of days ago she was getting ready to leave for work together with her husband Dustin. After they both left the house, he ran to catch his bus, and met up with Courtney, an acquaintance who took the same bus with him every day.	Laura works as a software engineer at a large company. **She likes her job, because she never has any business trips.** A couple of days ago she was getting ready to leave for work together with her husband Dustin. After they both left the house, he ran to catch his bus, and met up with Courtney, an acquaintance who took the same bus with him every day.	Dustin said to Courtney: “Laura was just getting ready for work with me. She grabbed her house keys!”
doctor	Bruce goes to his local medical practice every few years. Yesterday after leaving the practice he ran into his friend Sarah on the street, and they stopped to catch up. After they parted, Sarah walked on and soon saw Bruce's brother Drake on the street. She stopped to say Hi.	Bruce goes to his local medical practice every few years. Yesterday after leaving the practice he **was going, as always, to take the bus home, and** ran into his friend Sarah walked on the street. They stopped to catch up. After they parted, Sarah walked on and she soon saw Bruce's brother Drake on the street. She stopped to say Hi.	Sarah said to Drake: “Bruce was just leaving the medical practice. He got examined by the doctor!”

Table A.12 (*continued*)

Story	Context in Exp 1	Context in Exp-s 2 and 3	IR-utterance
shampoo	Olivia has beautiful hair, and pays a lot of attention to it. Today, when she was leaving the bathroom after showering, she ran into her roommate and best friend Thomas. She talked to him briefly about her hair, as she tends to do. Later that day, when their housemate Jill came home, she and Thomas started talking about Olivia.	Olivia has beautiful hair. **She pays a lot of attention to it, and never dyes her hair.** Today, when she was leaving the bathroom after showering, she ran into her roommate and best friend Thomas. She talked to him briefly about her hair, as she tends to do. Later that day, when their housemate Jill came home, she and Thomas started talking about Olivia.	Thomas said to Jill: “Olivia was talking to me about washing her hair. She used shampoo!”
skydiving	Jared takes skydiving courses at the local airfield, when he has free time. Last week he was at the skydiving center, with his friend Stella in the same group as him. They spent the day together, and when Stella went home in the evening, she texted Jared's brother Don, who was also a good friend of hers.	Jared takes skydiving courses at the local airfield, when he has free time. Last week he was at the skydiving center, with his friend Stella in the same group as him. **As always,** they spent the day together, and when Stella went home in the evening, she texted Jared’s brother Don, who was also a good friend of hers.	Stella said to Don: “Jared was in the skydiving course today. He jumped out of the plane!”
letter	Amy enjoys writing letters to people she is close to, especially around holidays. About two days ago, she wrote a letter to her cousin Michelle, and today she talked about it with her brother Steve. In the evening, Steve got a call from Michelle, and they started talking about family.	Amy enjoys writing letters to people she is close to, especially around holidays. **She always uses fountain pens for writing letters, as it is something deep and old fashioned.** About two days ago, she wrote a letter to her cousin Michelle, and today she talked about it with her brother Steve. In the evening, Steve got a call from Michelle, and they started talking about family. Steve said to Michelle: Amy wrote you a letter. She mailed it!	Steve said to Michelle: “Amy wrote you a letter. She mailed it!”
bus	Adam usually takes the bus to work, as the stop is a few blocks from his house. Last week, after he got off the bus, he ran into Virginia, his ex-girlfriend. They stopped for a little while to catch up.	**Adam lives quite far from his office.** Last week**, as always,** he took the bus to work, as the stop is a few blocks from his house. After he got off the bus, he ran into Virginia, his ex-girlfriend. They stopped for a little while to catch up.	Adam said to Virginia: “I took the bus this morning. I walked to the bus stop!”
clothes	Esther often goes along with her friends when they go clothes shopping, as it’s something she also enjoys. Today, when she was walking out of a mall after spending time with her friends, she ran into George, another old friend of hers. They decided to catch up while walking to the bus stop.	Esther often goes along with her friends when they go clothes shopping, as it’s something she also enjoys. Today, when she was walking out of the mall after spending time with her friends, she **bought an ice-cream, as she always does, and** ran into George, another old friend of hers. They decided to catch up while walking to the bus stop.	Esther said to George: “I was out clothes shopping. I tried something on!”
airplane	Greg frequently travels by air, to see family and attend conferences. Last week he flew to a conference, and met up with Helen, an old colleague he occasionally traveled with. They went to breakfast together, and started talking about their travel.	Greg frequently travels by air, to see family and attend conferences**, although he never spends his miles, when booking a ticket**. Last week he flew to a conference, and met up with Helen, an old colleague he occasionally traveled with. They went to breakfast together, and started talking about their travel.	Greg said to Helen: “I flew here. I took my cell phone on board with me!”
hair	Sandy usually cuts her own hair, although she has no training. Two days ago, after she gave herself another haircut, she went for a walk along her street. She quickly ran into her ex, Patrick, and they stopped to catch up for a few minutes.	Sandy usually cuts her own hair, although she has no **formal** training **and never dyes her hair at home**. Two days ago, after she gave herself another haircut, she went for a walk along her street. She quickly ran into her ex, Patrick, and they stopped to catch up for a few minutes.	Sandy said to Patrick: “I just cut my hair. I used scissors!”
exhibit	Henry often goes to art exhibitions, as there's an art museum a short walk from his place. Last week, after going to a new photography exhibition, he encountered his friend Max on his way home. They paused on the street and chatted for a while.	Henry often goes to art exhibitions, as there's an art museum a short walk from his place. Last week, after going to a new photography exhibition, he encountered his friend Max on his way home. **They almost missed each other, since Henry always listens to music on headphones when walking on the streets and at first he did not hear Max calling him out. Finally** they paused on the street and chatted for a while.	Henry said to Max: I just went to the new photo exhibit. “I looked at the photographs!”
drive	Helen works hard at her job, and enjoys the challenges she's given at work. Today, after driving her car to work as usual, she ran into her office-mate Peter while walking into the building. They stopped briefly to say hello.	Helen works hard at her job. She enjoys the challenges she's given at work **and never leaves the office before her boss.** Today, after driving her car to work as usual, she ran into her office-mate Peter while walking into the building. They stopped briefly to say hello.	Helen said to Peter: “I just parked my car. I locked it!”
pizza	Gary often orders pizza at work, from a famous pizzeria nearby. A few days ago, after he placed an order, his colleague Stephanie walked over to his cubicle to chat.	Gary often orders pizza at work, from a famous pizzeria nearby**, as he always gets some discount there**. A few days ago, after he placed an order, his colleague Stephanie walked over to his cubicle to chat.	Gary said to Stephanie: “I just ordered pizza. I picked the toppings!”

(*continued on next page*)

Table A.12 (*continued*)

Story	Context in Exp 1	Context in Exp-s 2 and 3	IR-utterance
dishes	Julia always tries to wash the dishes after eating, to avoid annoying her roommates. A few days ago, she was getting ready to go out after doing the dishes. She ran into her roommate Justin on her way out, and started talking to him.	Julia always tries to wash the dishes after eating, to avoid annoying her roommates**, as they always complain if there are some dishes left in the sink**. A few days ago, she was getting ready to go out after doing the dishes. She ran into her roommate Justin on her way out, and started talking to him.	Julia said to Justin: “I just did the dishes. I rinsed them!”
library	Emma often borrows books from the library, as she doesn't have much spare cash to spend. Last week, after going to the library, she was heading home with several books, and ran into her best friend Tim on the street. They stopped to quickly say hello.	no changes	Emma said to Tim: “I just got some books at the library. I checked them out!”

Table A.13
Corresponding questions in Experiments 1, 2, and 3.

Story	Target Question	Comprehension question in Exp 1	Comprehension question in Exp-s 2 and 3	Filler question in Exp-s 1, 2, and 3	Additional filler question in Exp-s 2 and 3
grocery	How often do you think John usually pays the cashier, when going shopping?	Where did John meet his roommate Susan?	How often do you think John's flatmate Susan does yoga at home?	How often do you think John usually gets apples, when going shopping?	How often do you think Susan and Peter usually talk to each other?
restaurant	How often do you think Mary usually eats, when going to a restaurant?	What is Mary's occupation?	How often do you think there is live music, in the restaurant where Mary went yesterday?	How often do you think Mary usually gets to see the kitchen, when going to a restaurant?	How often do you think Sally and David usually run into each other?
feed dog	How often do you think Jim usually throws the can away, when feeding the dog?	What pet does Jim have to feed?	How often do you think Jim feeds the dog in the evenings?	How often do you think Jim usually adds some medicine to the food, when feeding the dog?	How often do you think Carl and Lucy usually chat?
subway	How often do you think Jane usually buy a ticket, when taking the subway?	Where did Jane meet her friend Don today?	How often do you think Jane takes the subway to get around the city?	How often do you think Jane usually comes close to falling off the platform, when taking the subway?	How often do you think Beth and Don usually meet?
swimming	How often do you think Lisa usually brings her swimsuit, when going swimming?	What does Lisa like to do after work?	How often do you think the swimming pool nearby Lisa's office is open, when she finishes working?	How often do you think Lisa usually brings her children, when going swimming?	How often do you think Jen and Harvey usually see each other at the pool?
train	How often do you think Brian usually gets off at his stop, when taking the train?	Where did Brian meet his colleague Rachel?	How often do you think Brian buys a cup of coffee in the mornings, before getting on the train?	How often do you think Brian usually gets to work late, when taking the train?	How often do you think Oliver and Rachel usually see each other?
work	How often do you think Laura usually grabs her house keys, when getting ready for work in the morning?	What is Laura's occupation?	How often do you think Laura has business trips at her job?	How often do you think Laura usually puts on several layers of clothing, when getting ready for work in the morning?	How often do you think Dustin and Courtney usually talk to each other?
doctor	How often do you think Bruce usually gets examined by the doctor, when going to the medical practice?	Where did Bruce go yesterday?	How often do you think Bruce takes the bus home after going to his local medical practice?	How often do you think Bruce usually gets fitted with a heart rate monitor, when going to the medical practice?	How often do you think Drake and Sarah usually run into each other?
shampoo	How often do you think Olivia usually uses shampoo, when washing her hair?	What did Olivia talk about with her best friend Thomas?	How often do you think Olivia dyes her hair?	How often do you think Olivia usually finds some split ends, when washing her hair?	How often do you think Thomas and Jill usually chat with each other?
skydiving	How often do you think Jared usually jumps out of a plane, when going skydiving?	Where does Jared take skydiving courses?	How often do you think Jared and his friend Stella spend the day together, when going to a skydiving center?	How often do you think Jared is usually the first to jump, when going skydiving?	How often do you think Stella and Don usually talk?
letter	How often do you think Amy usually mails a letter, after writing it?	What does Amy enjoy doing?	How often do you think Amy uses simple rollerball pens for writing letters to people she is close to?	How often do you think Amy usually writes letters?	How often do you think Steve and Michelle usually talk to each other?
bus	How often do you think Adam usually walks to the bus stop, when	How does Adam usually get to work?	How often do you think Adam takes the bus to work?	How often do you think Adam usually barely has	How often do you think Adam and Virginia usually run into each other?

Table A.13 (*continued*)

Story	Target Question	Comprehension question in Exp 1	Comprehension question in Exp-s 2 and 3	Filler question in Exp-s 1, 2, and 3	Additional filler question in Exp-s 2 and 3
	taking the bus in the morning?			room to stand, when taking the bus in the morning?	
clothes	How often do you think Esther usually tries something on, when going clothes shopping?	Where did Esther meet her friend George?	How often do you think Esther buys an ice-cream after spending time with her friends in the mall?	How often do you think Esther usually comes across a big sale, when going clothes shopping?	How often do you think Esther and George usually see each other?
airplane	How often do you think Greg usually carries his cell phone on board with him, when flying on a plane?	Where did Greg fly last week?	How often do you think Greg spends his miles, when booking an airplane ticket?	How often do you think Greg usually gets into business class, when flying on a plane?	How often do you think Greg and Helen usually meet up?
hair	How often do you think Sandy usually uses scissors, when cutting her hair?	Who does usually cut Sandy's hair?	How often do you think Sandy dyes her hair at home?	How often do you think Sandy usually cuts her hair a bit shorter than intended, when cutting it?	How often do you think Sandy and Patrick usually see each other?
exhibit	How often do you think Henry usually looks at photographs, when going to a photo exhibit?	What exhibition did Henry attend last week?	How often do you think Henry listens to music on headphones, when walking on the streets?	How often do you think Henry usually decides to buy a photograph, when going to a photo exhibit?	How often do you think Henry and Max usually talk to each other?
drive	How often do you think Helen usually locks her car, after parking it?	How does Helen usually get to work?	How often do you think Helen leaves the office before her boss?	How often do you think Helen usually discovers that one of her tail lights has gone out, when parking her car?	How often do you think Helen and Peter usually run into each other?
pizza	How often do you think Gary usually picks the toppings, when ordering pizza?	What does Gary often order at work?	How often do you think Gary gets a discount, when ordering pizza from the famous pizzeria nearby?	How often do you think Gary usually orders pizza at work?	How often do you think Gary and Stephanie usually talk?
dishes	How often do you think Julia usually rinses the dishes, when doing them?	What does Julia always try to do after eating?	How often do you think Julia's flatmates complain, if there are some dishes left in the sink?	How often do you think Julia usually polishes the dishes, when doing them?	How often do you think Julia and Justin usually see each other?
library	How often do you think Emma usually checks out the books, when getting some books at the library?	Where was Emma coming from when she met her best friend Tim?	How often do you think Emma buys books?	How often do you think Emma usually looks at the library's exhibit, when getting some books at the library?	How often do you think Emma and Tim usually run into each other?

Appendix B. Distributions in the main and the secondary tasks

Experiment 1

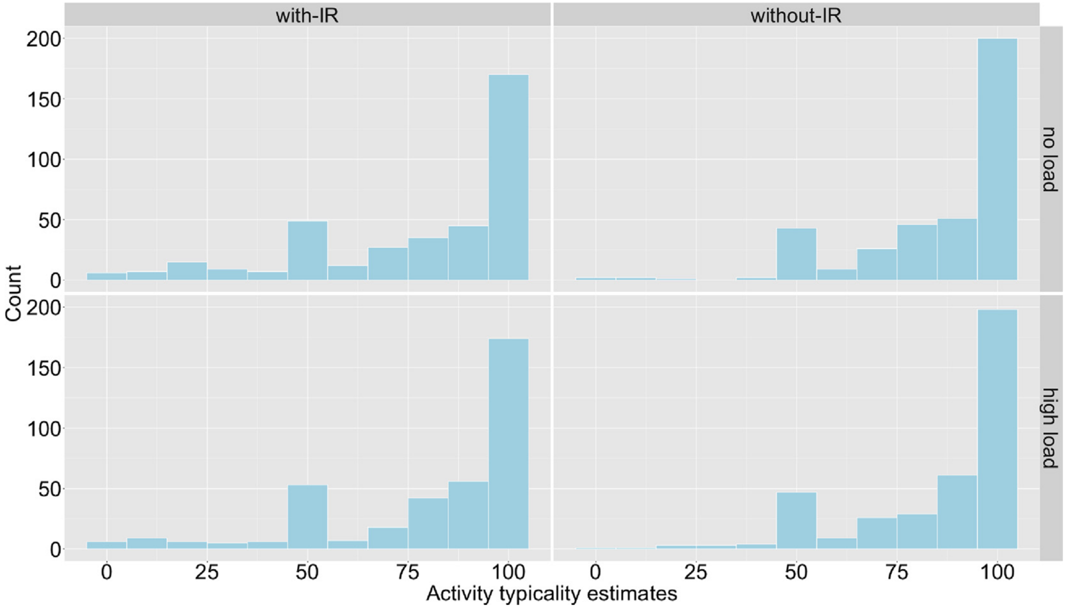

Fig. B.14. Experiment 1. Distribution of the non-transformed ratings in the target question by story (with-IR vs. without-IR) and cognitive load (high vs. no).

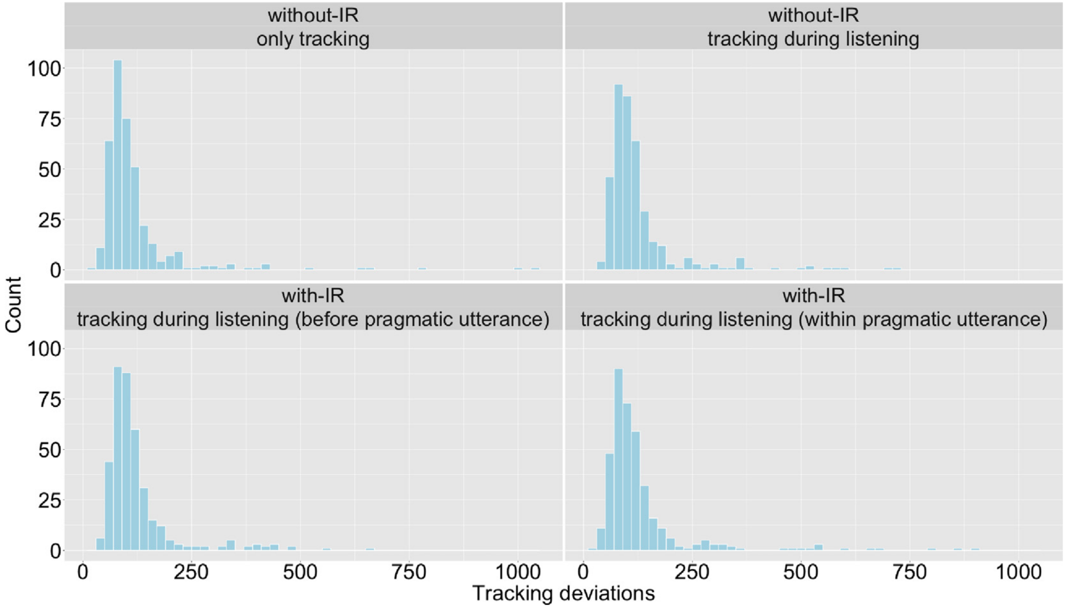

Fig. B.15. Experiment 1. Distribution of the non-transformed by-subject mean tracking deviations in the four intervals of interest.

Experiment 2

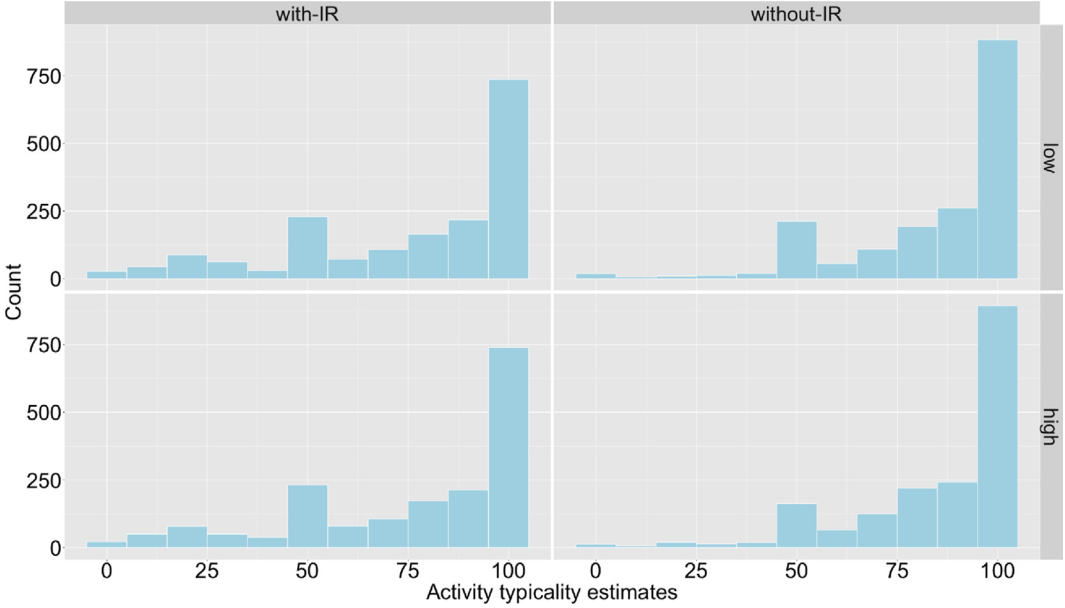

Fig. B.16. Experiment 2. Distribution of the non-transformed ratings in the target question by story (with-IR vs. without-IR) and cognitive load (high vs. low).

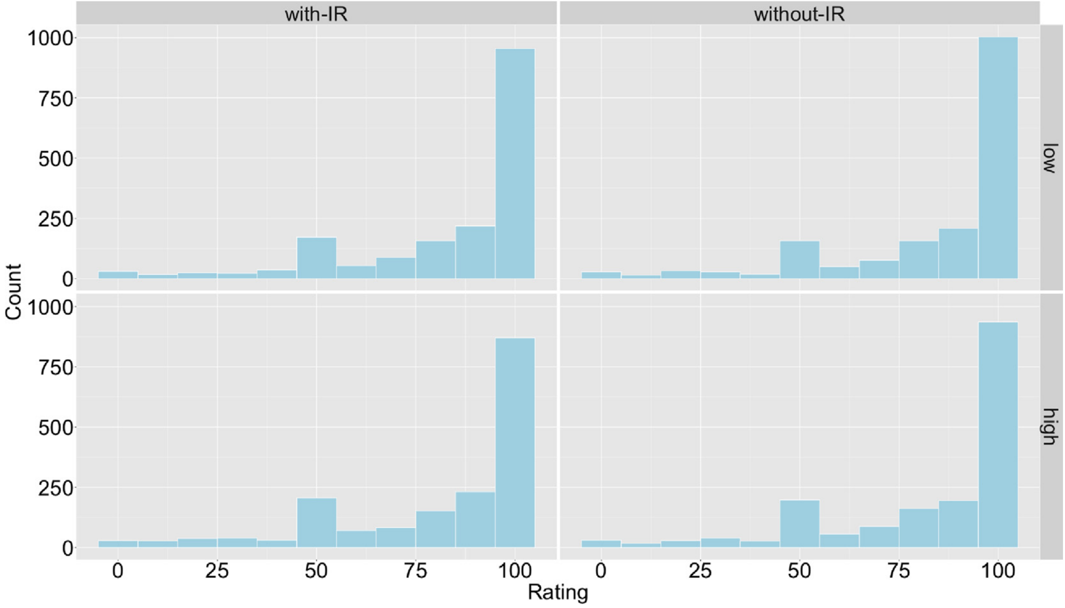

Fig. B.17. Experiment 2. Distribution of the ratings in the comprehension question (the higher the rating, the more correct the response is) by story (with-IR vs. without-IR) and cognitive load (high vs. low) conditions.

Appendix C. Bayesian analysis of target linguistic judgements

Experiment 1

Table C.14
Experiment 1. Prior specification for the Bayesian model of target activity typicality ratings.

prior	class	coef	group
student_t (3, 0, 2.5)	Intercept		
(flat)	b	story	
(flat)	b	load	
(flat)	b	story:load	
student_t (3, 0, 2.5)	sd	Intercept	item
student_t (3, 0, 2.5)	sd	story	item
student_t (3, 0, 2.5)	sd	Intercept	subject
student_t (3, 0, 2.5)	sd	story	subject
student_t (3, 0, 2.5)	sd	load	subject
lkj_corr_cholesky (1)	L		item
lkj_corr_cholesky (1)	L		subject
gamma (0.01, 0.01)	phi		

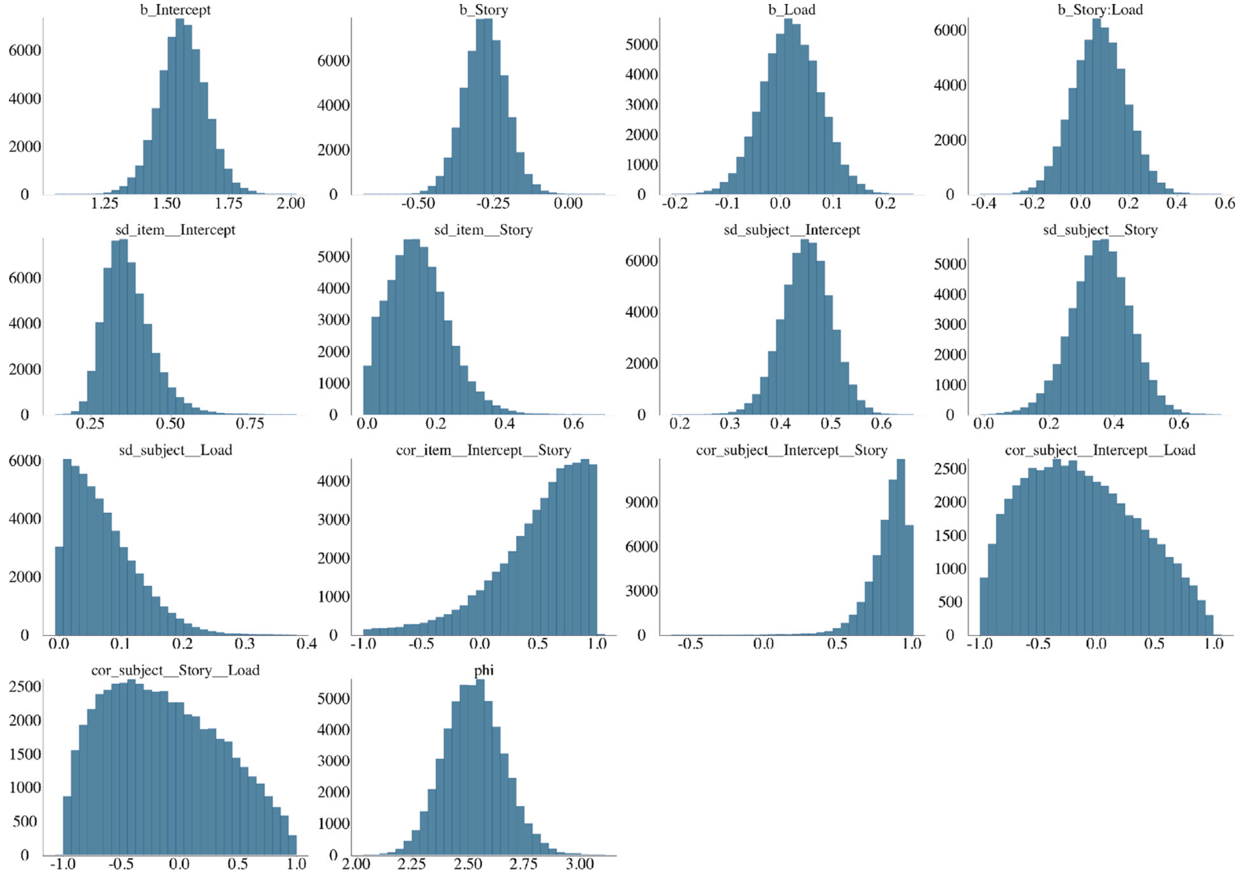

Fig. C.18. Experiment 1. Posterior distributions of parameter estimates in the hierarchical beta regression model of transformed typicality ratings.

Experiment 2

Table C.15
Experiment 2. Prior specifications for the Bayesian model of target activity typicality ratings.

default	non-informative	weakly informative	more informative	class	coef	group
student_t (3, 0, 2.5)	normal (0, 10)	normal (1.56, 1)	normal (1.56, 0.5)	Intercept		
(flat)	normal (0, 10)	normal (-0.28, 0.5)	normal (-0.28, 0.28)	b	story	
(flat)	normal (0, 10)	normal (0, 2)	normal (0, 0.2)	b	load	
(flat)	normal (0, 10)	normal (0, 2)	normal (0, 0.4)	b	story:load	
student_t (3, 0, 2.5)	as default	as default	as default	sd	Intercept	item
student_t (3, 0, 2.5)	as default	as default	as default	sd	story	item
student_t (3, 0, 2.5)	as default	as default	as default	sd	Intercept	subject
lkj_corr_cholesky (1)	as default	as default	as default	L		item
gamma (0.01, 0.01)	normal (1, 10)	normal (2.5, 2)	normal (2.5, 2)	phi		

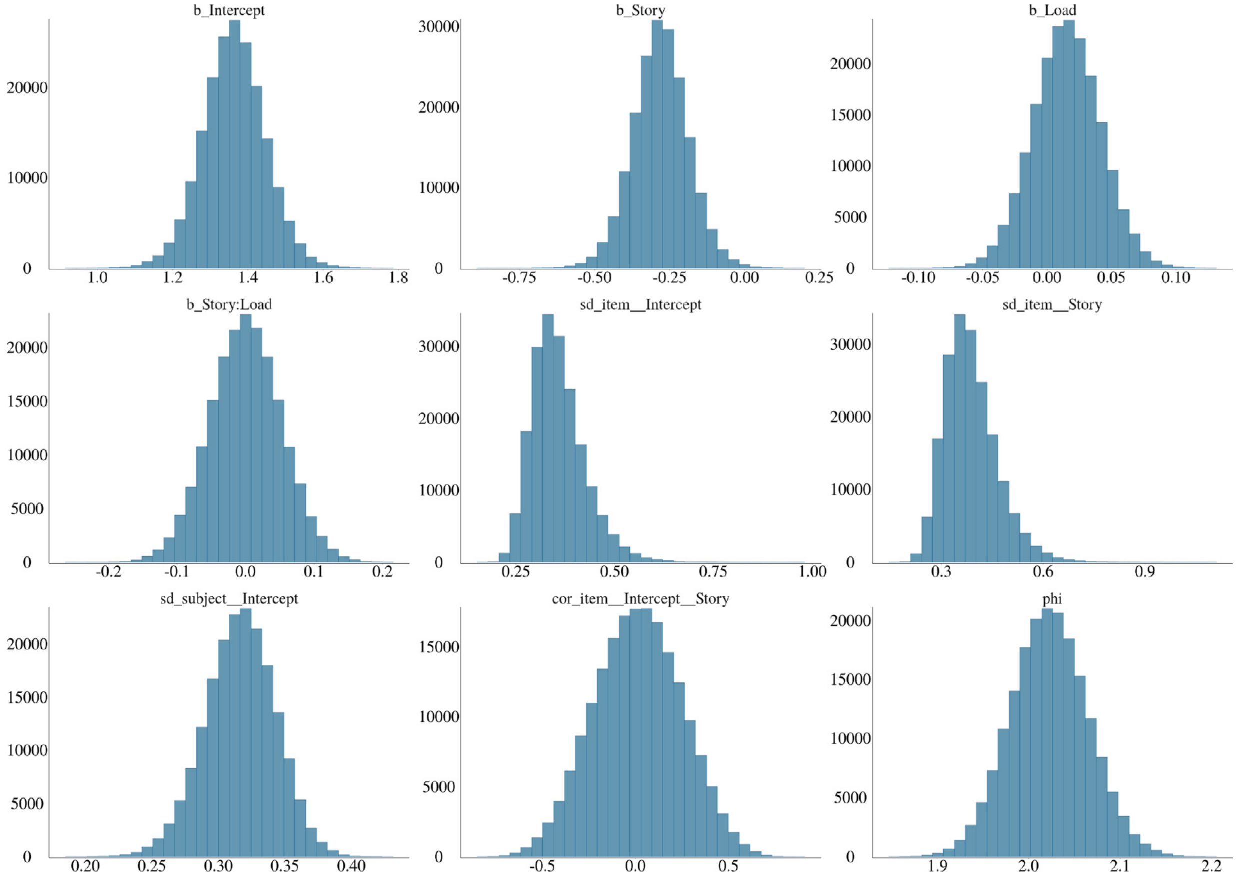

Fig. C.19. Experiment 2. Posterior distributions of parameter estimates in the hierarchical beta regression model of transformed typicality ratings.

Appendix D. Power analysis

The results of the first experiment put us in a good position for conducting a meaningful power analysis for the second experiment. In particular, we are interested in how many participants we have to collect in order to have a good chance of detecting an interaction effect. For the power analysis, we set the target power level $\beta = 80\%$ and the target significance level $\alpha = 0.05$. In order to determine the number of participants, we also need to set a target effect size which we want to be able to detect. However, previous studies largely do not report effect size. We therefore decided to proceed as follows: we set our target effect size for the interaction to half of the size of the main effect. This means that we want to make sure that we can detect a pragmatic effect reduction such that the typicality rating in the high load with-IR condition is reduced by half as much as the typicality rating in the low load with-IR condition compared to the baseline without-IR condition.

We then decided to take a two-step approach. As a first rough approximation, we estimate the order of magnitude for the number of subjects based on experiment 1. As there are some differences with respect to the experimental design between experiments 1 and 2, we then use the resulting sample for a more precise power calculation. As power calculation libraries don't currently exist for linear mixed effects models with the beta family, we instead performed our power calculation using a normal distribution. Note that the results of the power analysis should be treated as conservative estimates of the lower bound of power, since the typicality ratings are strongly negatively skewed and the beta family gives a better fit to the data, and has more power to detect an effect than regression using a Gaussian distribution.

We follow recommendations by Gelman (2018) in running a power analysis for the main effect of story (with-IR vs. without-IR) based on experiment 1. The model included story and load as fixed effects, but used a gaussian distribution. The random structure included by-subject and by-item intercepts and by-item random slope for story – see model output in Table D.16. Power analysis was performed in R 3.6.3 using ‘simr’ 1.0.5 package (Green and MacLeod, 2016). The analysis showed that a power of 80% for the main effect of story can be reached with 200 subjects ([80.42; 85.18] 95% confidence interval for β) which, based on (Gelman, 2018)'s connection between main and interaction effects, suggested that for an interaction that is half the size of the main effect, one would need more than 1000 subjects. The power curve for a story predictor and details of power analysis are shown in Figure D.20.

Subsequently, we calculated a more exact power estimation based on the data from 770 subjects collected in the experiment 2 setup. We again ran a linear mixed effects regression model which included story, load, and their interaction as fixed effects. The random structure included by-subject and by-item intercepts and by-item random slopes for story. The main effect of story was equal to $\beta = -7.93$ (t-value $= -4$, p-value $< .001$ – see full model output in Table D.17). In experiment 2, we planned to hire not less than 1400 subjects. Assuming this number of subjects, we ran power analysis in ‘simr’ package to estimate the power to detect an effect of interaction that is half of the size of the main effect of story. Based on 2000 simulations, we found that the power to detect such an effect was estimated as 100% with a 95% confidence interval of [99.82%, 100%] given 1400 subjects. For the actual study, we decided to collect approximately 30% participants more.

Table D.16
Experiment 1. Effect sizes (b), standard errors (SE), t-values, and p-values in the linear mixed effects regression model of the target activity typicality ratings. The model was used in the power estimation of the main effect of story – see Figure D.20. Significance codes: *** 0.001 | ** 0.01 | * 0.05.

	b	SE	t	p
Intercept	82.48	1.66	49.72	***
Story: with-IR	−6.4	1.46	−4.39	***
Load: high	0.8	1.	0.8.	ns
Random Effects	*Variance*			
Subject	105.44			
Item	44.58			
Story|Item	22.91			

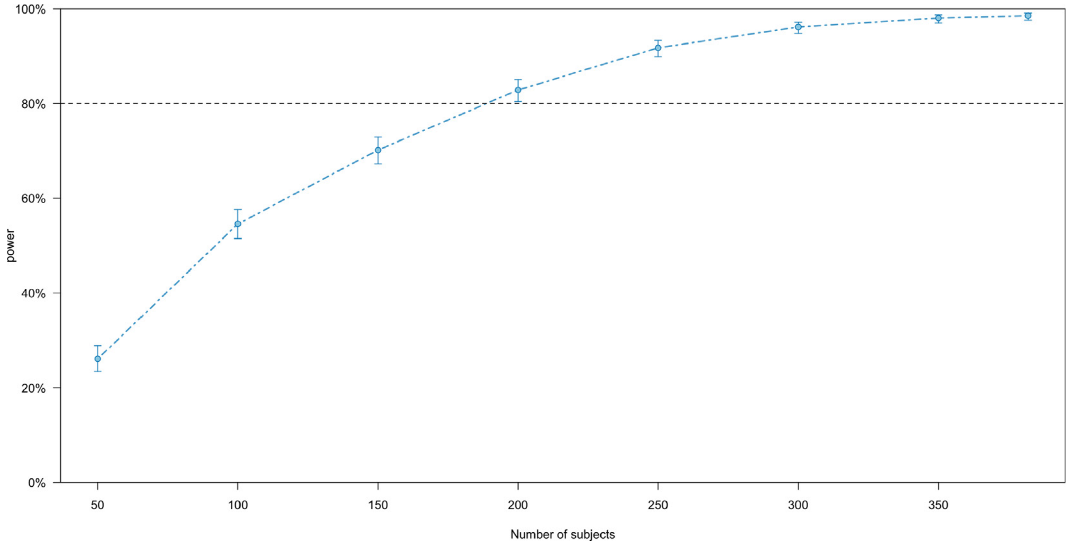

Fig. D.20. Observed power (±95% CI) to detect a fixed effect of story with size '-6.4' calculated over a range of sample sizes. For a power estimation, a likelihood ratio test was used with parameter $\alpha = 0.05$ (1000 iterations).

Table D.17
Experiment 1. Effect sizes (b), standard errors (SE), t-values, and p-values in the linear mixed effects regression model of the target activity typicality ratings. The model was used to estimate the number of subjects needed to detect a significant effect of interaction twice smaller than the main effect of load with the power of 80%. Significance codes: *** 0.001 | ** 0.01 | * 0.05.

	b	SE	t	p
Intercept	80.51	1.8	44.82	***
Story: with-IR	−7.93	1.98	−4	***
Load: high	0.53	0.77	−0.69	ns
Story*Load: with-IR	−0.44	1.53	−0.28	ns
Random Effects	*Variance*			
Subject	83.92			
Item	59.46			
Story|Item	67.09			

References

Antoniou, K., Cummins, C., Katsos, N., 2016. Why only some adults reject under-informative utterances. J. Pragmat. 99, 78–95.

Baddeley, A.D., Bressi, S., Della Sala, S., Logie, R., Spinnler, H., 1991. The decline of working memory in alzheimer's disease: a longitudinal study. Brain 114, 2521–2542.

Baddeley, A.d., Eysenck, M., 2010. Anderson. Memory mc (2009).

Baker, R., Gill, A., Cassell, J., 2008. Reactive redundancy and listener comprehension in direction-giving. In: Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, pp. 37–45.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. 68, 255–278.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67, 1–48. https://doi.org/10.18637/jss.v067.i01.

Bott, L., Bailey, T.M., Grodner, D., 2012. Distinguishing speed from accuracy in scalar implicatures. J. Mem. Lang. 66, 123–142.

Bott, L., Noveck, I.A., 2004. Some utterances are underinformative: the onset and time course of scalar inferences. J. Mem. Lang. 51, 437–457.

Bower, G.H., Black, J.B., Turner, T.J., 1979. Scripts in memory for text. Cognit. Psychol. 11, 177–220.

Bürkner, P.C., 2017. brms: an R package for Bayesian multilevel models using Stan. J. Stat. Software 80, 1–28. https://doi.org/10.18637/jss.v080.i01.

Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Maechler, M., Bolker, B.M., 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R Journal 9, 378–400. URL: https://journal.r-project.org/archive/2017/RJ-2017-066/index.html.

Brown, P.M., Dell, G.S., 1987. Adapting production to comprehension: the explicit mention of instruments. Cognit. Psychol. 19, 441–472.

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

Cho, J., 2020. Memory load effect in the real-time processing of scalar implicatures. J. Psycholinguist. Res. 1–20.

Chwilla, D.J., Kolk, H.H., 2005. Accessing world knowledge: evidence from n400 and reaction time priming. Cognit. Brain Res. 25, 589–606.

De Neys, W., Schaeken, W., 2007. When people are more logical under cognitive load: dual task impact on scalar implicature. Exp. Psychol. 54, 128–133.

Degen, J., Tanenhaus, M.K., 2015. Processing scalar implicature: a constraint-based approach. Cognit. Sci. 39, 667–710.

Degen, J., Tanenhaus, M.K., 2019. Constraint-based pragmatic processing. The Oxford Handbook of Experimental Semantics and Pragmatics 21–38.

Degen, J., Tessler, M.H., Goodman, N.D., 2015. Wonky worlds: listeners revise world knowledge when utterances are odd. In: CogSci.

Demberg, V., Sayeed, A., 2016. The frequency of rapid pupil dilations as a measure of linguistic processing difficulty. PLoS One 11, e0146194.

Dieussaert, K., Verkerk, S., Gillard, E., Schaeken, W., 2011. Some effort for some: further evidence that scalar implicatures are effortful. Q. J. Exp. Psychol. 64, 2352–2367.

Engonopulos, N., Sayeed, A., Demberg, V., 2013. Language and cognitive load in a dual task environment. In: Proceedings of the Annual Meeting of the Cognitive Science Society.

Fairchild, S., Mathis, A., Papafragou, A., 2020. Pragmatics and social meaning: understanding under-informativeness in native and non-native speakers. Cognition 200, 104171.

Fairchild, S., Papafragou, A., 2021. The role of executive function and theory of mind in pragmatic computations. Cognit. Sci. 45, e12938.

Fairs, A., Bögels, S., Meyer, A.S., 2018. Dual-tasking with simple linguistic tasks: evidence for serial processing. Acta Psychol. 191, 131–148.

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, 121.

Foppolo, F., Mazzaggio, G., Panzeri, F., Surian, L., 2021. Scalar and ad-hoc pragmatic inferences in children: guess which one is easier. J. Child Lang. 48, 350–372.

Gelman, A., 2018. You need 16 times the sample size to estimate an interaction than to estimate a main effect. Statistical Modeling, Causal Inference, and Social Science. [blog post]

Geurts, B., 2010. Quantity Implicatures. Cambridge University Press.

Green, P., MacLeod, C.J., 2016. Simr: an r package for power analysis of generalized linear mixed models by simulation. Methods Ecol. Evol. 7, 493–498.

Grice, H.P., 1975. Logic and conversation. In: Speech Acts. Brill, pp. 41–58.

Grigoroglou, M., Papafragou, A., 2019. Children's (and adults') production adjustments to generic and particular listener needs. Cognit. Sci. 43, e12790.

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, 42–55.

Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M., 2004. Integration of word meaning and world knowledge in language comprehension. Science 304, 438–441.

Holding, D.H., 1989. Counting backward during chess move choice. Bull. Psychonomic Soc. 27, 421–424.

Horn, L., 1993. Economy and redundancy in a dualistic model of natural language. Sky 1993, 33–72.

Horn, L., 2014. Information structure and the landscape of (non-) at-issue meaning. In: The Oxford Handbook of Information Structure.

Horn, L.R., 1972. On the Semantic Properties of Logical Operators in English. University of California, Los Angeles.

Horn, L.R., 1991. Given as new: when redundant affirmation isn't. J. Pragmat. 15, 313–336.

Huang, Y., 2012. The Oxford Dictionary of Pragmatics. Oxford University Press.

Kravtchenko, E., Demberg, V., 2015. Semantically underinformative utterances trigger pragmatic inferences. In: CogSci.

Kravtchenko, E., Demberg, V., 2022. Informationally redundant utterances elicit pragmatic inferences. Cognition 225, 105159.

Kursat, L., Degen, J., Denison, S., Mack, M., Xu, Y., Armstrong, B.C., 2020. Probability and processing speed of scalar inferences is context-dependent. In: CogSci.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. Lmertest package: tests in linear mixed effects models. J. Stat. Software 82.

Levinson, S.C., 2000. Presumptive Meanings: the Theory of Generalized Conversational Implicature. MIT press.

Li, K.Z., Lindenberger, U., Freund, A.M., Baltes, P.B., 2001. Walking while memorizing: age-related differences in compensatory behavior. Psychol. Sci. 12, 230–237.

Lockridge, C.B., Brennan, S.E., 2002. Addressees' needs influence speakers' early syntactic choices. Psychonomic Bull. Rev. 9, 550–557.

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.

Mazzarella, D., 2014. Is inference necessary to pragmatics? Belg. J. Linguist. 28, 71–95.

Noveck, I.A., 2001. When children are more logical than adults: experimental investigations of scalar implicature. Cognition 78, 165–188.

Plummer, P., Eskes, G., 2015. Measuring treatment effects on dual-task performance: a framework for research and clinical practice. Front. Hum. Neurosci. 9, 225.

Pype, A., Lin, J., Murray, S., Boynton, G., 2010. Individual differences in the shape of visual attention during object tracking. J. Vis. 10, 315–315.

Recanati, F., 2004. Literal Meaning. Cambridge University Press.

Regneri, M., Koller, A., Pinkal, M., 2010. Learning script knowledge with web experiments. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 979–988.

Rogowsky, B.A., Calhoun, B.M., Tallal, P., 2016. In: Does Modality Matter? The Effects of Reading, Listening, and Dual Modality on Comprehension, vol. 6. Sage Open, 2158244016669550.

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., Kaufmann, S., 2012. Communicating with cost-based implicature: a game-theoretic approach to ambiguity. In: The 16th Workshop on the Semantics and Pragmatics of Dialogue, Paris, September.

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., Iverson, G., 2009. Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bull. Rev. 16, 225–237.

Ryzhova, M., Mayn, A., Demberg, V., 2022. In: What Inferences Do People Actually Make on Encountering a Redundant Utterance? An Individual Differences Study (unpublished Manuscript). https://www.uni-saarland.de/fileadmin/upload/lehrstuhl/demberg/Ryzhova2022whatinferences.pdf.

Schad, D.J., Nicenboim, B., Bürkner, P.C., Betancourt, M., Vasishth, S., 2021. Workflow techniques for the robust use of bayes factors. arXiv preprint arXiv: 2103.08744.

Schank, R.C., Abelson, R.P., 1975. Scripts, Plans, and Knowledge. IJCAI, pp. 151–157.

Scholman, M.C., Demberg, V., Sanders, T.J., 2020. Individual differences in expecting coherence relations: exploring the variability in sensitivity to contextual signals in discourse. Discourse Process 57, 844–861.

Sedivy, J.C., 2007. Implicature during real time conversation: a view from language processing research. Philos. Compass 2, 475–496.

Teresa Guasti, M., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., Meroni, L., 2005. Why children and adults sometimes (but not always) compute implicatures. Lang. Cognit. Process. 20, 667–696.

Vogels, J., Demberg, V., Kray, J., 2018. The index of cognitive activity as a measure of cognitive processing load in dual task settings. Front. Psychol. 2276.

Vogels, J., Howcroft, D.M., Tourtouri, E., Demberg, V., 2020. How speakers adapt object descriptions to listeners under load. Lang. Cognit. Neurosci. 35, 78–92.

Walker, M.A., 1993. Informational Redundancy and Resource Bounds in Dialogue. Graduate School of Arts and Sciences, University of Pennsylvania. Ph.D. thesis.

Wilson, D., Sperber, D., 2002. Relevance Theory.

Yang, X., Minai, U., Fiorentino, R., 2018. Context-sensitivity and individual differences in the derivation of scalar implicature. Front. Psychol. 1720.

Zufferey, S., Moeschler, J., Reboul, A., 2019. Implicatures. Cambridge University Press.

Margarita Ryzhova is a PhD student at Saarland University, chair of Computer Science and Computational Linguistics, Germany. Her research focuses on the discourse processing, cost and individual differences in pragmatic inferences.

Vera Demberg is a Professor of Computer Science and Computational Linguistics at Saarland University, Germany. Her area of expertise includes experimental and computational discourse and pragmatics, experimental psycholinguistics, and language processing in a dual-task setting. Her research has been published in journals such as Cognition; Journal of Memory and Language; Frontiers in Psychology; Discourse Processes; Language, Cognition and Neuroscience. She is currently an editor for the journal "Dialogue and Discourse".