Informationally redundant utterances elicit pragmatic inferences

Ekaterina Kravtchenko , Vera Demberg [*]

Saarland University, Campus C7.2, Saarland Informatics Campus, 66123 Saarbrücken, Germany

ABSTRACT

Most theories of pragmatics and language processing predict that speakers avoid excessive informational redundancy. Informationally redundant utterances are, however, quite common in natural dialogue. From a comprehension standpoint, it remains unclear how comprehenders interpret these utterances, and whether they make attempts to reconcile the 'dips' in informational utility with expectations of 'appropriate' or 'rational' speaker informativity. We show that informationally redundant (overinformative) utterances can trigger pragmatic inferences that increase utterance utility in line with comprehender expectations. In a series of three studies, we look at utterances which refer to stereotyped event sequences describing common activities (*scripts*). When comprehenders encounter utterances describing events that can be easily inferred from prior context, they interpret them as signifying that the event conveys new, unstated information (i.e. an event otherwise assumed to be habitual, such as *paying the cashier* when *shopping*, is reinterpreted as non-habitual). We call these inferences *atypicality inferences*. Further, we show that the degree to which these atypicality inferences are triggered depends on the framing of the utterance. In the absence of an exclamation mark or a discourse marker indicating the speaker's specific intent to communicate the given information, such inferences are far less likely to arise. Overall, the results demonstrate that excessive conceptual redundancy leads to comprehenders revising the conversational common ground, in an effort to accommodate unexpected dips in informational utility.

1. Introduction

Informationally redundant utterances are predicted to be infelicitous by pragmatic theories and theories of language processing (cf. Aylett & Turk, 2004, for a theory of phonetic reduction; Cohen, 1978, for a computational theory of speech act generation; Grice, 1975, for a theory of rational communication; or Jaeger, 2010, for a theory of reduction at all levels of linguistic representation, among many others). In this article, we explore in a series of studies whether informational redundancy can elicit pragmatic inferences.

So what do we mean by redundancy? At the *form level*, redundancy may include overt mention of, or increased articulatory effort towards producing material that is easily predictable or recoverable in context. In other words, more signal is provided than the comprehender requires to accurately recover the intended phonological, lexical, or syntactic form. Examples of redundancy avoidance at this level include vowel shortening (Aylett & Turk, 2004), use of shorter word variants (Mahowald, Fedorenko, Piantadosi, & Gibson, 2013), or omission of optional complementizers (Jaeger, 2010). At the *informational or conceptual level*, redundancy refers to the explicit mention of information that the

comprehender is already in a position to infer automatically, using world knowledge or common ground information, or that is already entailed or strongly implied by the preceding discourse. In other words, more information is provided than needed to recover the intended meaning or world state. In contrast to redundancy avoidance at the form level, constraints against *overinformativeness*, or redundancy, at the informational level have always been somewhat debated (Grice, 1975).

There is ample evidence that speakers are routinely overinformative at the informational level, and that speaker overinformativity is frequently tolerated by listeners (Baker, Gill, & Cassell, 2008; Engelhardt, Bailey, & Ferreira, 2006; Nadig & Sedivy, 2002; Walker, 1993). In this paper, we explore the question of whether there is empirical evidence for comprehenders noticing and reacting to informational redudancy by deriving pragmatic inferences. Specifically, we look at cases where the redundancy is at the level of background world knowledge – as opposed to, for example, repeating something that has already been stated, or referring to an object in more descriptive detail than strictly necessary given the physical context.

Consider for instance utterance (1), which is at face value redundant, in that it overtly states that "John" *paid the cashier*, which conventionally

can be inferred from context simply on the basis of common sense knowledge about everyday activities (Bower, Black, & Turner, 1979; cf. Zwaan, Magliano, & Graesser, 1995), here: him having gone *shopping*.

(1) "John went grocery shopping. He paid the cashier!"

A theoretic account of utterance choice which places a constraint on informational redundancy would predict that uttering the second sentence in this context would be marked, at best. Further, it would predict that comprehenders should note this markedness, and might react to it by drawing pragmatic inferences. In this paper we show that comprehenders, through pragmatic reasoning about the common ground, can accommodate these utterances by changing their previous beliefs about the likely world state. In the case of the concrete example, the previous world state is the belief that people usually pay when shopping. When the informationally redundant utterances is present, the listeners infer that this statement is informative in the case of John, and hence infer that John does not usually pay the cashier. We term such inferences that accommodate the redundant utterance by revising beliefs about the habituality of an event as *atypicality inferences*.

1.1. World knowledge

Utterances like the one shown in (1) are redundant on the basis of background world knowledge. As background knowledge is fairly unsystematic and comprehender-specific, and can be difficult to control for, here we use *script*, or *schema* knowledge, which constitutes a specific type of world knowledge. *Script* knowledge refers to people's implicit awareness of the typical event structures of various stereotyped activities, such as *going shopping* or *going to a restaurant* (Fillmore, 2006; Minsky, 1975; Schank & Abelson, 1977). The former, for example, normally involves events such as *going to a store*, *selecting food items*, and *paying the cashier*. Comprehenders anticipate upcoming events once a script is "invoked" (Zwaan et al., 1995); and when recalling stories based on scripts, have difficulty distinguishing actions that were actually mentioned, and those that were unmentioned but implied by the script (Bower et al., 1979). These findings suggest that events which are strongly associated with a script are almost part of its conventional meaning, and that explicitly mentioning their occurrence is therefore redundant.[1]

Utterance (1) introduces a well-known script or event sequence (*grocery shopping*), followed by an informationally redundant event description (*he paid the cashier!*), which references a highly predictable sub-event from the script. In this example, the event described in the second sentence is already strongly implied to have occurred by the preceding invocation of the *grocery shopping* script – given the assumption, shared by most speakers and comprehenders, that people overwhelmingly pay cashiers when they go grocery shopping. Mentioning it explicitly, therefore, is redundant.

1.2. Informational redundancy

First, we want to address a problem of terminology. In most experimental work, informational redundancy has been described as a problem of *overinformativeness*, *overspecification* or *overdescription*, and as addressed by the second part of Grice's Quantity Maxim, which states that speakers should provide no more information than is necessary to get their message across. However, the term *informativeness* in the

pragmatic literature has been used to refer to both informational redundancy (Engelhardt et al., 2006; Grice, 1975) and its converse, as well as to the relative informativeness of terms in an implicational scale (Horn, 1984; Levinson, 2000). The latter variety of informativeness, now more typically associated with the Quantity Maxim, is invoked more in reference to unjustified vagueness where a more precise description is available, but where both descriptions are similar or equal in length and effort. When referring to informational redundancy or its converse, in contrast, the issue is more one of either excessive or insufficient wordiness – for example, as in the case of overinformative nominal modification (such as using *the big red cup* or *the cup on the towel* to identify the only available cup in a given context), where speakers might choose to describe objects in more detail than is strictly necessary. In this paper we concern ourselves strictly with overinformativeness in the sense of informational redundancy, as originally described by Grice (1975), and in the literature on nominal overspecification.

While most pragmatic theories do address cases where a speaker may be informationally redundant (Grice, 1975; Horn, 1984; Levinson, 2000, among many others), they often leave open the question of whether comprehenders do, in fact, perceive (unjustified) redundancy as infelicitous, as well as how they interpret redundant utterances. Most accounts do argue that comprehenders expect speakers to behave rationally – namely, by communicating in a way that is consistent with getting across the intended message (which, furthermore, should be truthful). However, as Grice (1975) notes, it is unclear whether excessive redundancy comes into any real conflict with the goal of successful (truthful, sufficiently informative, relevant, etc.) communication – although comprehenders may wonder what the "point" of excessive information is, and attempt to rationalize unexpected "dips" in informational utility by infusing them with additional pragmatic meaning. Informationally redundant utterances do not clearly interfere with comprehension, as *underinformativeness* or underspecification does, and may aid comprehension in some cases (e.g., object identification; cf. Nadig & Sedivy, 2002; Rubio-Fernández, 2016).[2] In this light, it is not straightforwardly clear whether overinformativeness constitutes non-rational speaker behavior, and specifically to what degree this part of the *Quantity* maxim holds: do not make your contribution more informative than is required."[3]

It is, however, possible that comprehenders perceive excessive information as, at minimum, non-relevant to the discourse (Grice, 1975; Horn, 1984). The question, then, is whether comprehenders make any particular note of redundancy, simply find it odd or infelicitous, or attempt to accommodate it. If comprehenders do perceive redundant information as irrelevant, then rational speakers should avoid overtly stating conceptually redundant information, except in those cases where this information is intended to communicate a more informative non-literal meaning (or signal an unusual world state). Correspondingly, comprehenders where possible ought to interpret conceptually redundant utterances as either an attempt to convey some non-literal (relevant and informative) meaning, or as reflecting a background world state where the information conveyed can't be taken for granted, and is therefore informative. How comprehenders do in fact react to redundancy has to date only been empirically investigated within the relatively narrow scope of nominal modification in referent identification tasks.

These tasks typically instruct participants to look at or somehow engage with items such as: *the [red] apple*, *the [tall] boot* (Davies &

---

[1] Highly inferable events are occasionally used as temporal anchors (*After she entered the restaurant, she*…), and may be used to transition back from interruptions to the script (*She stopped to talk to Brad on the street. She then entered the restaurant*…). However, outside of these contexts, easily inferable script events are usually not mentioned overtly (Bower et al., 1979; Regneri, Koller, & Pinkal, 2010).

[2] This is not to say that comprehension is not in any way impaired by redundancy, and in fact we suspect that it is - but at face value, there is nothing about receiving more information than needed that necessarily hinders one from arriving at the intended meaning of a message.

[3] Grice (1975) explicitly suspected that it did not; later accounts separated the notions of semantic vs. form-based informativeness, which is discussed in the following section.

Katsos, 2010, 2013; Engelhardt et al., 2006; Nadig & Sedivy, 2002; Pogue, Kurumada, & Tanenhaus, 2016; Sedivy, 2003). What has been found is that in interactive, spontaneous speech, speakers frequently modify nouns with adjectives that are not strictly necessary for referent identification (e.g., referring to a cup as *the red cup*, in a context where there are no other cups of any color) (Engelhardt et al., 2006; Nadig & Sedivy, 2002). Studies also showed that overinformative descriptions are often easily tolerated by comprehenders and can in fact be helpful for comprehension, when they describe non-canonical properties, or properties which may speed up object identification (Engelhardt et al., 2006; Rubio-Fernández (2016); Pogue et al., 2016; Arts, Maes, Noordman, & Jansen, 2011; Rehrig, Cullimore, Henderson, & Ferreira, 2021; Long, Rohde, & Rubio-Fernandez, 2020; Mangold & Pobel, 1988; Paraboni, Van Deemter, & Masthoff, 2007; Paraboni & van Deemter, 2014; Sonnenschein & Whitehurst, 1982; Tourtouri, Delogu, Sikos, & Crocker, 2019). There is, however, also evidence that informationally redundant utterances which have no apparent (e.g., perceptual) utility are unlikely to be produced, are generally judged to be relatively infelicitous, and tend to generate inferences (Davies & Katsos, 2010; Sedivy, 2003). As Rohde, Futrell, and Lucas (2021) have shown, utterances with highly predictable or unsurprising content can even lead to longer reading times compared to utterances with less predictable content. In their study, they use a context manipulation that raises expectations for something unusual: "My cousin Mary is a surprising person who never does things the way you'd expect." They show that in such a context, the target "shovel" in "For instance, in order to chop some carrots, she was using a shovel yesterday in the afternoon." has shorter reading times than the more locally predictable tool "knife".

More generally, there is still some difficulty in distinguishing what constitutes informational redundancy, which creates difficulty in determining the precise theoretical implications of previous work (e.g., perceptually helpful "redundant" adjectives are questionably redundant in the first place, in the sense of having communicative utility). Additionally, previous studies are limited by the fact that they uniformly focus on a very particular, and relatively concise variety of informational redundancy, which is further bound to a specific class of lexical items, raising the question of to what degree it is possible to generalize from the results. This points towards a need to look at informational redundancy in the context of utterances and constructions that are both quite costly for speakers, and have no readily apparent utility to comprehenders - either in terms of perception or comprehension, or in terms of facilitating the completion of a task. Further, we would argue that it is important to investigate constructions that are less bound to a specific set of lexical items, and are more likely to be perceived as flouting of a conversational norm against redundancy - for example, complex and lengthy multi-word utterances such as those in Example (2).

Ultimately, the question of how comprehenders treat overinformativeness is relevant to a more general theory of human communication, and should be answered to determine the extent to which: a) comprehenders and/or speakers consistently behave in a "Gricean" manner; b) under which conditions they do so, and which deviations from communicative norms are more likely to occur/be tolerated; and c) to what extent comprehenders attempt to resolve apparent violations, and which strategies they use to do so. If comprehenders do not appear to make much of overinformativeness (whether in terms of inferences made, or maxim violations perceived), and there is little evidence that speakers deliberately use overinformative utterances to convey specific non-literal meanings, then it is questionable to what degree overinformativeness violates communicative norms, in the first place.

1.3. Pragmatic inferences about the common ground

To date there has been relatively little work on the different strategies comprehenders might employ in making sense of an apparent violation of conversational maxims. Most work has focused on the

scenario where a comprehender detects an apparent maxim violation, assumes that the speaker is in fact being cooperative, and comes up with an additional, non-literal meaning that the speaker may have intended (which repairs the apparent violation). Another strategy is simply to assume that the speaker is being plainly uncooperative. A third strategy, which has received little attention, is that of modifying background assumptions about the world in which events take place, if doing so would repair the apparent violation.

Among the few works addressing the modification of background assumptions in depth is Degen, Tessler, and Goodman (2015), who investigated comprehenders' willingness to revise their assumptions about the assumed common ground, in response to utterances whose pragmatic meaning would otherwise be inconsistent with it. They found that background assumptions about the world are surprisingly defeasible: comprehenders frequently accommodate the pragmatic meaning of utterances such as "*some of the marbles sank*" (upon being thrown into a pool), by assuming that the utterances signify a strange scenario where physics doesn't quite work as expected. Similarly, in our work here, we explore inferences related to a shift in common ground assumptions. In the case of Example (1), a speaker states that *John*, having gone shopping, *paid the cashier*. A comprehender might then "repair" the redundancy by not extending the generalization that people usually pay the cashier when going shopping to *John*, thus inferring that *John* does not in fact habitually pay the cashier. The experiments conducted as part of this article explore the willingness of comprehenders to shift background assumptions in different contexts.

Research on conventionalized inferences, and specifically scalar implicatures, has been critical to developing formal linguistic theory, due to the role they play in disambiguating pragmatic and semantic contributions to utterance meaning. However, context-dependent (ad-hoc) inferences, which occur far more frequently and ubiquitously, are similarly important to developing a more general theory of human communication (as originally intended by Grice, 1975). The body of experimental work teasing apart which properties of utterances trigger, alter, or modulate the strength of pragmatic inferences is still relatively small – however, having a more comprehensive model of cues which are taken into account by comprehenders, when interpreting utterances, is necessary both for building models of pragmatic reasoning, and for interpreting empirical results. In addition, there is a general need for further quantitative data on the specific conditions under which inferences are generated, in order to develop and test predictions of formal models of pragmatic reasoning (cf. Frank & Goodman, 2012).

1.4. Hypotheses of the present study

In the present article, we conduct a series of experiments which expose readers to informationally redundant utterances and assess to what extent these redundant utterances lead to pragmatic inferences about the common ground. In this section, we lay out three alternative hypotheses regarding how comprehenders respond to the informational redundancy. Specifically, we will consider what might happen when a comprehender encounters one of our experimental utterances (which are embedded within a larger context in the experiment):

(2) John just came back from the grocery store. **He paid the cashier.**

Hypothesis 1. No inference.

The first possibility is that comprehenders do not find informational redundancy particularly marked, as it does not necessarily interfere with interpreting the intended message – or, at most, find redundant utterances slightly odd or suboptimal, as has been found in some studies (Davies & Katsos, 2010). In the case of our utterance (2), in this scenario, we might expect that comprehenders would interpret the utterance literally, and make no more of it than stated; i.e., they would take away

the message that on some particular instance, John paid the cashier, and perhaps the speaker described it in a bit more detail than strictly necessary.

Hypothesis 2.   Non-detachability.

If comprehenders do expect speaker utterances to always have a certain level of informational utility, then they may attempt to resolve the provision of excessive or unnecessary information by drawing pragmatic inferences, regarding what they think the utterance means or signifies from the speaker's perspective. These pragmatic inferences would then serve to increase the informational utility of the utterance, and allow comprehenders to maintain the belief that the speaker is being cooperative – since assigning an "informative" pragmatic meaning to an apparently redundant utterance in effect removes the redundancy. In the case of utterance (2), comprehenders might conclude that John's *cashier-paying* is being announced due to its being unusual or unexpected, and that John can't therefore typically be counted on to pay the cashier. This reaction should occur as long as the background and linguistic context is basically consistent with that interpretation, and, as in the case of most pragmatic inferences, should be unaffected by changes to the utterance which do not alter its semantic content (generally referred to as *non-detachability*; Grice, 1975), such as prosodic and/or discourse markers which do not change the truth conditions of the sentence – i.e., the inference should be attached to the semantic content, not the specific linguistic form of the utterance.

Hypothesis 3.   Form sensitivity.

The third possibility is that, as in Hypothesis 2, comprehenders react to a statement of *John's* having paid the cashier by inferring that *John* must be a habitual non-payer. However, as the inferences we are concerned with are based, in a sense, on the specific *form* of the utterance (i. e., too much signal is used to communicate something that would have already been understood), it is possible that such inferences may be relatively sensitive to how exactly the utterance is expressed.[4] In particular, we suspect that expending extra articulatory effort on expressing our already redundant utterance would increase the strength of any pragmatic inferences drawn (or even cause inferences to be drawn where none would be otherwise). In the case of our utterance, what we would predict in this case is that the more obvious effort is expended on producing the utterance (whether in the form of prosodic stress, or another attention-drawing signal of relevance and intentionality), the stronger the inference. To note, some amount of form sensitivity is not necessarily incompatible with the second hypothesis (*non-detachability*), but the complete absence of an inference would be.

In this paper, we present three experiments, run concurrently on the same population, which test whether informationally redundant event descriptions give rise to pragmatic inferences.[5] The first experiment uses implicit exclamatory prosody (the marker "*!*") to signal that the utterance is an intentionally conveyed, important, and relevant piece of information. The second experiment uses the discourse marker "*oh yeah, and…*" to do the same, while avoiding the surprise conventionally implied by the exclamation mark. In the third experiment, we predict that informational redundancy by itself, in absence of prosodic or discourse cues as to relevance and intentionality, triggers weaker,

consistent with the third hypothesis (*form sensitivity*).

2.   Experiment 1: Implicit intent signaled by an exclamation mark

Experiment 1 tests whether informationally redundant event descriptions trigger *atypicality* inferences when the utterance is apparently effortful, intentional, and attentionally prominent – here signaled by an exclamation mark at the end of the utterance (this would disprove the "no inference" hypothesis). Exclamatory intonation is a natural way of introducing something that may be noteworthy or unusual (Rett, 2011), without otherwise altering the semantic content of the utterance.

We present naive participants with a limited number of brief narratives, which set up the common ground context, and the relationships between discourse participants. The narratives then include a brief dialogue which contains the informationally-redundant target utterance (or a non-redundant control). After reading the narratives, participants rate how habitual they believe certain behaviors in the story to be with respect to the character in the story. We expect that for activities with high habituality (e.g. paying when going shopping), comprehenders generally expect that these activities will be followed also by the characters in the story (i.e., that John also usually pays when shopping). In the case of a mention of an informationally redundant utterance, we however expect that comprehenders may revise this default inference, and instead infer that the mentioned event may not be highly habitual for the person mentioned in the story (i.e., that John does not usually pay). In contrast, those participants who are not exposed to the informationally redundant utterance are expected to maintain the default habituality assumptions (e.g., John pays when shopping).

2.1.   Methods

2.1.1.   Participants

700 eligible participants (760 total; median age bracket 26–35; 50% female) completed the experiment; they were recruited on Amazon Mechanical Turk. The target number of eligible participants was predetermined through a simulation power analysis (adapted from Arnold, Hogan, Colford, & Hubbard, 2011): all predicted higher-order interactions, assuming effect sizes determined by the results of the pilot experiments from Kravtchenko and Demberg (2015) were detectable at > 0.80. The R code and a plot for the power analysis can be found in the supplementary materials.[6]

The task was open only to workers located in the US, and with an approval rating of ≥ 95%. All workers were asked to state their native childhood language (with no penalty for stating a language other than English, to encourage accurate reporting), age bracket (under 18, 18–25, 26–25, and up, in intervals of 10), and gender. Those who did not indicate English, or listed their age as outside the interval of 18–65, were excluded from all analysis (60; 7.89%), with additional participants recruited to replace them.

Prior to seeing any experimental items, participants were given three practice questions with a total of 11 sliders,[7] unrelated to the experimental stimuli, which used continuous sliding scales ranging from *Never* to *Always* (or similar), like in the experiment, see Fig. 1. Unlike the experimental stimuli, these questions had 'correct' answers – such as *How likely is a fair coin to come up heads twice, if flipped 10 times? (very unlikely–very likely)*. If participants provided responses that could not be

---

[4] The principle of *non-detachability* generally does not hold for Manner implicatures (Grice, 1975), which are conceptually similar to the inferences we look at here; for space reasons, we will not go into more detail on this type of implicature, however.

[5] A pilot experiment was originally published at the CogSci conference as Kravtchenko and Demberg (2015). The experiments reported in this article differ from the original pilot in that they have an increased sample size, and were conducted concurrently on the same general population, to ensure that their results could be compared directly. The stimuli were also redesigned to read more naturally, and fillers were included.

---

[6] https://osf.io/84jdp/?view_only=ff5859d3f33b485d95254395f95a52dc

[7] The first question asked participants to move the slider completely to the right (first slider), and completely to the left (second slider); the second question asked about what kind of pet John most likely got yesterday, with one slider each for "dog", "whale", "cat", "alligator"; the final question asked about the likelihood of the number of heads (0–4) for flipping a coin 4 times, with one slider for each possible answer.

Never	Sometimes	Always

Fig. 1. This is a slider, as used by experiment participants.

judged reasonably accurate, they were asked to re-read the instructions, and respond again.

Those who did still not provide accurate or plausible responses to the trial questions were unable to proceed to the main task, and their data as a result was not recorded by the platform (e.g., those who rated the likelihood of 50% heads on multiple fair coin flips as low, compared to other possible outcomes).[8] Participants were likewise unable to proceed in the study, or submit their results, without having answered all questions. These participant were hence not counted into the number of 700 participants that completed the study.

2.1.2. Design

Our experiment uses a 2 (context: ordinary vs. wonky) by 3 (activity type: conventionally habitual vs. non-habitual vs. no activity) within-participants design. Our critical condition is the ordinary context followed by a conventionally habitual event. The "no activity" condition means that the critical utterance is omitted. This condition is a baseline that allows us to calculate general beliefs about the habituality of the target event, when it is not explicitly mentioned. We term the estimates that we obtain from this condition the by-item *pre-utterance beliefs* (see Section 2.1.4 below for more details). The wonky context and the non-habitual activity serve as control conditions to our critical condition.

The primary question of interest is whether informationally redundant utterances (in this case, descriptions of highly *habitual* activities) trigger pragmatic inferences. Specifically, we test for inferences concerning of the revision of common-ground beliefs about the *habituality* of the informationally redundant mention of the habitual activity. Consider the following example:

(3) CONVENTIONALLY HABITUAL EVENT:
   "John just came back from the grocery store. **He paid the cashier!**"

The bolded utterance here, given a "default" or *ordinary* common ground, is ***informationally redundant***. We hypothesize that readers will notice the informational redundancy, and try to accommodate it by revising the common ground belief about the habituality of paying the cashier. The reader will hence infer that *John* does *not* habitually pay the cashier, as such a scenario would justify the overt mention of *John's* cashier-paying. The informational redundancy arises due to the high *conceptual predictability* of *paying the cashier*, and is resolved if one assumes that this activity is not as habitual, or predictable as initially assumed. We will refer to this condition as the *conventionally habitual* activity.

One of our control conditions checks whether the inference (that an activity is less habitual than would otherwise be expected) can be cancelled by manipulating the common ground.

The activity described becomes "non-habitual" given a *wonky* common ground[9] such as in (4), where the context suggests that typical assumptions (e.g., that some given individual would *pay the cashier* when they *go to the grocery store*) may not hold. At that point, the activity description ceases to be informationally redundant, and the inference should therefore not arise. This control condition keeps the description

itself constant and manipulates only the common ground. It thus ensures that any effect we measure is in fact due to the presence of informational redundancy, and verifies that comprehenders are sensitive to discourse context.

(4) WONKY COMMON GROUND CONTEXT: John habitually doesn't pay.
   "John just came back from the grocery store. **He paid the cashier!**"

Finally, we wanted to provide a baseline for "typical" interpretation of non-redundant event descriptions; and to confirm that similarly structured descriptions of *non-habitual* activities, as in (5), do not provoke similar inferences (which would suggest a problem with the stimulus design or response measure). In (5), the utterance is not informationally redundant, and is not expected to generate any specific inferences. We also wanted to confirm that the *wonky* common ground in the previous example does not significantly affect the interpretation of conventionally *non-habitual* event mentions (which would suggest that there is an unexpected effect of context manipulation on stimulus interpretation, in general):

(5) NON-HABITUAL EVENT:
   "John just came back from the grocery store. **He got some apples!**"

As in (4), participants should draw no atypicality inferences here, as the event described is not (typically) overly habitual. These conditions therefore provide a secondary control measure.

2.1.3. Materials

24 stimuli were constructed as brief stories/narratives, based on distinct stereotyped scripts or events. Each story had one of 2 context types (*ordinary* vs. *wonky* common ground, relative to the *conventionally habitual* script activity). In all stories, declarative utterances, spoken by one of the discourse participants, described one of 2 types of script activities (*conventionally habitual* vs. *non-habitual*), making a total of 4 conditions, plus two conditions for collecting pre-utterance beliefs, see below. An example of the stimulus material with the different conditions is shown in (1).

(1) EXPERIMENTAL STIMULI

[1a] John **often goes to the grocery store around the corner from his apartment**ordinary	[1b] John **is typically broke, and doesn't usually pay when he goes to the grocery store**wonky
[2] 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.	
[3] Susan said to Peter:"John just came back from the grocery store.	
[4a] He **paid the cashier**conv. habitual**!**"	[4b] He **got some apples**non-habitual**!**"

Participants saw stories which consisted of either context version 1a or 1b, the story content in 2 and 3, as well as one of the critical utterances 4a or 4b. In addition, we also collected data for short versions of the stories that end after text segment 2, in order to collect estimates of how habitual activities are believed to be based on the context alone (*pre-utterance beliefs*). We correspondingly denote the beliefs collected based on the whole stories as *post-utterance beliefs*. All materials were presented in written form. Note though that the effects reported here

---

[8] Since this data was not recorded, we cannot report on the number of participants who were unable to proceed to the main task.

[9] We borrow the term *wonky* from Degen et al. (2015), where it is similarly used to describe non-default common grounds, in which typical rules as to how things proceed are expected to not hold, and which comprehenders may assume when encountering otherwise pragmatically infelicitous utterances.

have also been replicated using spoken stimuli (Ryzhova & Demberg, 2020). The complete list of stimuli can be found in our online repository: https://osf.io/h5afr/?view_only=ff5859d3f33b485d95254395f95a52dc.

Following each passage, participants were queried as to how habitual they believed the *conventionally habitual* and *non-habitual* activities (as well as 2 other scenario-relevant distractor activities) were, for the person who was the subject of the discourse (the individual mentioned in the context 1a or 1b):

Q1 How often do you think John usually *pays the cashier*, when going shopping?

Q2 How often do you think John usually *gets apples*, when going shopping?

Q3 How often do you think John usually goes to the grocery store?

Q4 How often do you think Susan and Peter usually talk to each other?

Each question could be responded to on a continuous sliding scale of "Never" to "Always" (see Fig. 1). The slider itself was not visible until the participant clicked on the point on the scale that they thought was most appropriate, to avoid having people default towards a particular value. After they responded to all questions, participants could submit their answers. Once they did, the next passage was displayed on a new screen.

Half of the stimuli included three discourse participants – one of whom engaged in the script activity (*John*), the second who learned from that participant that they engaged in it (*Susan*), and the third to whom the second communicated this fact (*Peter*). The other half only included two participants – the subject of the discourse, who engaged in the activity (*John*), and the second participant to whom they communicated this fact (*Susan*). Compared to the example above, for instance, *John* might instead be communicating directly to *Susan*: "*I just got back from the grocery store. I paid the cashier!*".

The construction of these stimuli was constrained in several ways. The scripts (e.g., *going shopping*) needed to be sufficiently complex to include multiple subactivities or subroutines, and there needed to be habitual as well as non-habitual subactivities (*paying the cashier, getting apples*). It needed to be possible for the script to play out without the habitual activity having taken place – otherwise, the discourse would be incoherent, or the inference would not be drawn. For example, one arguably cannot play *tennis* at all, without using a *racket*. There was also established common ground between all discourse participants, so that all were plausibly (from the point of view of the reader) aware of the typical habits of the discourse subject, particularly with regard to the activity described. Finally, the activities needed to be sufficiently stereotyped and (relatively) culturally invariant, so that participants could be expected to agree on what a script entailed, which activities were or weren't obligatory to the script sequence, etc.

All stimuli were normed on three qualities (in separate tasks): whether the activity fell into the *habitual* or *non-habitual* activity bin; whether the common ground manipulation was effective; and whether participants found it plausible that the script could be engaged in without the *habitual* activity. For activity predictability norming, participants were asked to rate the habituality of the activity (on a 0–100 scale), with an arbitrary cutoff of 70 between activity types. *Non-habitual* activities were on average rated 48.0 (25.1–68.1), and *habitual* activities were rated 87.8 (78.1–95.2). For common ground norming, participants rated *habitual* activities in *ordinary* (mean 83.4 [72.2–96.9]) or *wonky* common grounds (mean 39.2 [20.7–62.0]), with a within-item difference between the two of at least 15 points (mean difference 44.2; [19.8–72.9]); *non-habitual* activities had to score below 70 regardless of common ground (mean 45.2; on average 10.7 points higher in the *ordinary* common ground). For plausibility norming, a statement in the form of "*John **went shopping**, but didn't pay the cashier*" was rated as either *coherent* (plausible) or *incoherent* (implausible), with criteria being a majority of participants finding the statement coherent (*habitual*:

91% [67%–100%]; *non-habitual*: 94% [80%–100%]).

2.1.4. Measures

Our main measure of comprehender belief is derived from the slider positions that participants choose as an answer to the question about event habituality. The continuous response scale was discretised into numbers ranging from 0 (*Never*) to 100 (*Always*).

Pre-utterance beliefs, or baseline beliefs regarding activity habituality, were estimated from responses to stimuli presented without the activity description (see the next section for a more detailed explanation). The responses, aside from setting baseline measures (*pre-utterance beliefs*) of activity habituality, also provide an additional norming measure for how likely it is that a particular activity would be engaged in, in the context of a given script. Thus, activities which are more or less habitual, within a given class, can be compared against one another.

Post-utterance beliefs regarding activity habituality were estimated from responses to stimuli which included the redundant or non-redundant utterance (activity description), or where the activity description/utterance was visible.

Belief change due to reading the activity description was determined by modeling the magnitude and direction of difference between *pre-utterance beliefs* and *post-utterance beliefs*.

2.1.5. Procedure

The experiment was run interleavedly with experiments 2 and 3, which are described below. This was done to make sure that time and AMT population of workers were comparable. We ran the experiments in small rotating batches (of 9, or less): a batch of 9 participants completed the first experiment, after which the second experiment was scripted to go live until it was completed by 9 participants, and so forth. The only difference between the three experiments was the manipulation of full stop vs. exclamation mark vs. the discourse marker. Running them concurrently on the same population therefore allows us to directly compare their results. All workers who participated in an experiment were automatically disqualified from participating in any future batches; i.e., no participant took part in more than one experiment or batch. All experiments were run using the same interface.

Prior to seeing any experimental items, participants were given three practice questions to make sure they followed the instructions, see section 2.1.1. For the main experiment, participants were asked to read 6 experimental stimuli randomly selected out of the total of 24, as well as 4 filler items.[10] Each condition was only presented once, as follows: 2 of the stories were presented without the dialogue and event description (context and setting up of common ground only), and 4 stories were presented in their entirety (context, setting up of common ground, and the dialogue/event description). The 2 partial stories allowed us to collect measures of *pre-utterance beliefs* regarding activity habituality, and the 4 full stories gave us measures of *post-utterance beliefs* conditioned on the event description.

The experiment thus employed a between-subject design for belief measures, where *pre-utterance* and *post-utterance* belief estimates for any given item were provided by different participants, to eliminate the possibility of participants conditioning their *post-utterance* estimates not only on inferences made from the text, but also on their own *pre-utterance* estimates.[11] The 4 filler stimuli had the same structure as the critical items, but with the dialogue portion replaced by script-neutral utterances: "*You know, I'm really tired.*", "*Hey, do you know what time it is?*", "*So, what are you up to?*", or "*Have you heard the news today yet?*"

---

[10] To note, this means that each participant saw each manipulation only once, and the number of fillers was equal to the number of stimuli presented with dialogue.

[11] Note though that the results below largely mirror the results of a within-subjects version of the study reported in Kravtchenko and Demberg (2015).

2.2. Results

For the purposes of determining whether participants made any in-ferences regarding activity habituality, we modeled *belief change*, i.e. the difference between *pre-utterance* and *post-utterance* beliefs, or activity habituality estimates made with and without seeing the activity description. *Conventionally habitual* and *non-habitual* activities were modeled separately, as the *non-habitual* activity was used primarily as a control, and manipulations of common ground context did not otherwise target it. All factors were effect/sum coded. Note that the habituality estimates of the habitual and non-habitual activities are not easily comparable, as they are based on different questions (question 1 about the habituality of paying the cashier vs. question 2 about habitually buying apples). Running these conditions as separate analysis allows us to include in a single analysis all estimates that are based on the same question.

2.2.1. Conventionally habitual activities ("Paid the cashier")

The results for conventionally habitual activities are illustrated in Fig. 2, using violin plots. *Pre-utterance belief* ratings (obtained from participants who did not see the activity descriptions) showed that in *ordinary* context, these activities are indeed perceived as highly habitual (85.79 on a 0–100 scale). As predicted, *post-utterance belief* ratings for the condition where the conventionally habitual event is mentioned in the conversation, show lower habituality for the *ordinary context* activities (72.37) than *pre-utterance belief* ratings.

In the *wonky* context, the same activity is perceived as relatively non-habitual for John, with a priori ratings of 48. There was little change in participants' ratings when the conventionally habitual activity was mentioned (45.71) for *post-utterance beliefs*.

A linear mixed effects regression analysis, the results of which are summarized in Table 1, showed that the interaction between context and belief measure is statistically reliable ($\beta = -10.77$, $p < 0.001$). This interaction is driven by lowered activity habituality ratings when the readers see the utterance in a *ordinary* context ($\beta = -13.21$, $p < 0.001$).

In this experiment as well as the two following experiments, we used linear mixed effects models with the maximal random effects structure that was justified by the design. This means that we included by-subject random intercepts and slopes for common ground context (*ordinary / wonky*) and belief measure (*pre-utterance / post-utterance*), as well as by-item random intercepts and slopes for both factors and their interaction (Barr, Levy, Scheepers, & Tily, 2013). 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). *P*-values were obtained using the Satterthwaite approximation for degrees of freedom, as implemented in the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017).

These results show that, as predicted, when a *conventionally habitual* activity is explicitly described in a *ordinary* common ground context (i.e.
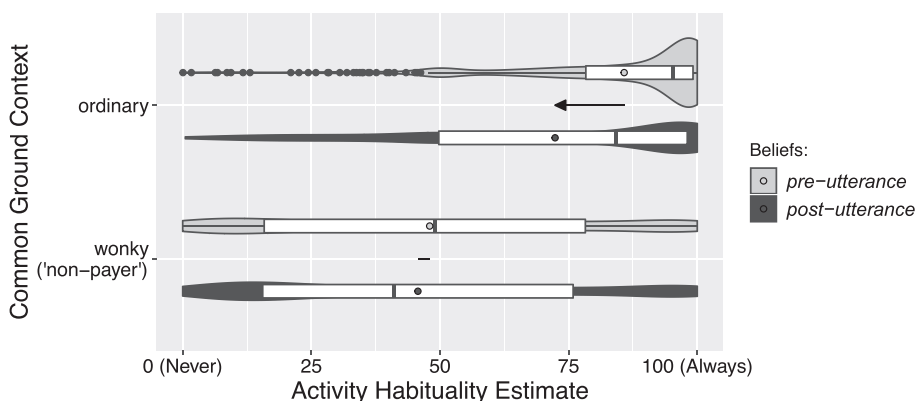
Table 1

Experiment 1: conventionally habitual (cashier-paying) activity analysis. This table shows the beta coefficients associated with each main effect in the model, as well as corresponding standard errors, t-values, and significance levels.

	β	SE(β)	t	p
Intercept	63.03	1.84	34.32	< 0.001
Common Ground: Ordinary	32.38	3.33	9.72	< 0.001
Belief: Post-utterance	−7.83	1.71	−4.58	< 0.001
Common Ground * Belief	−10.77	2.40	−4.50	< 0.001

a context in which the activity can be automatically inferred), many readers infer that the *conventionally habitual* activity must in fact be *non-habitual*; i.e., unusual for the individual who is the subject of the story, and therefore worth mentioning explicitly.

2.2.2. Non-habitual activities ("Bought some apples")

There was little change in participants' ratings of conventionally *non-habitual* activities from *pre-utterance beliefs* to *post-utterance beliefs* (*ordinary*: 40.80 to 42.47; *wonky*: 38.49 to 39.56), see Fig. 3.

A linear mixed effects regression analysis showed that estimates of activity habituality do not vary with the common ground context, nor are they conditioned on the utterance describing the activity (see Table 2). This is also consistent with our predictions, and indicates both that the context alteration does not inherently cause a change in activity habituality estimates (regardless of how script-central the activity is), and that conventionally *non-habitual* activities, given our *ordinary* context, are not interpreted as less habitual when mentioned.

2.3. Discussion

The results of the first experiment indicate that comprehenders do in fact perceive informational redundancy, in the form of mention of overly habitual activities, as a possible violation of conversational norms, and that they resolve this violation by reinterpreting the activity as the habituality not generalizing to the subject mentioned in the text. Comprehenders react to redundancy as they typically do to other apparent maxim violations – by assuming an implied non-literal meaning, or alternate background world state, that resolves the apparent violation. This runs in some contradiction to the initial ambivalence Grice (1975) expressed about the existence of such a constraint, and equivocal evidence from studies of informationally redundant nominal modification.

These results rule out the "no inference" hypothesis outlined in Section 1.4, and raise two questions that we address in the following experiments, regarding the importance of (implicit) prosody, and that the speaker signaling intentionality of the activity description. Our critical utterance used in experiment 1 was marked with an exclamation mark. We start out by noting that an exclamation mark may serve multiple purposes: it may signal surprise as to the course of described events, a speaker's intentionality in communicating a piece of

Fig. 2. Experiment 1: *conventionally habitual* (*cashier-paying*) activity analysis. This plot shows changes in activity habituality estimates depending on whether the utterance is seen, as well as whether the context causes the utterance activity to be perceived as non-habitual. Violin plots, overlaid with box plots, show the distribution of estimates. A violin plot is simply a smoothed and mirrored histogram: the fatter the distribution at a given point, the more instances there are of that particular activity habituality estimate. Circles represent mean values. Arrows show statistically significant differences between *before/pre-utterance* and *after/post-utterance* ratings.
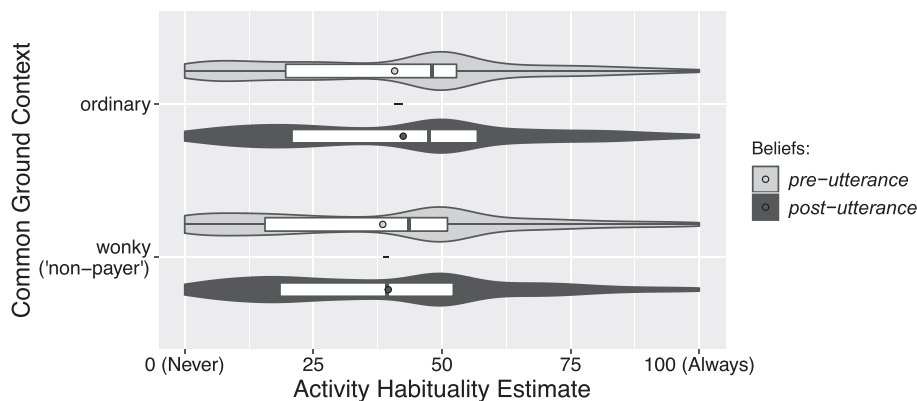
Fig. 3. Experiment 1: *non-habitual* (*apple-buying*) activity analysis.

Table 2
Experiment 1: non-habitual (apple-buying) activity analysis.

	β	SE(β)	T	p
Intercept	40.29	1.86	21.69	**< 0.001**
Common Ground: Ordinary	2.88	2.07	1.39	0.2
Belief: Post-utterance	1.34	1.85	0.73	0.5
Common Ground * Belief	0.01	2.14	0.01	1

information (i.e., the speaker displays clear and conscious intent to draw to the comprehender's attention the face that a given event occurred – as opposed to stalling for time, thinking of something to say, aborting a previously planned utterance, simply being uncooperative, and so forth), the importance and relevance of the information conveyed to the general discourse and comprehender's interests, and that the information preceding the exclamation point constitutes an "encapsulated" message in its own right, rather than serving as a temporal or causal anchor. (For example: *He paid the cashier. Then he noticed it was his classmate.*) Although it could be argued that the exclamation point (often a signal of surprise; Rett, 2011) forces a relative "non-habitual activity" interpretation independent of utterance informativity, this is not a likely explanation, as no signs of a similar effect are present in any of the other conditions.

Therefore, the first question is: how generalizable is the effect, and does the inference arise in contexts that do not implicitly signal the unexpectedness of the information conveyed (beyond the point that it is mentioned at all)? There is relatively little work on the question of which contextual cues specifically are employed by listeners in computing context-dependent inferences, as well as how these cues influence final interpretation. To test this, we use a discourse marker ("*Oh yeah, and…*"), which does not clearly signal surprise, in Experiment 2. This marker however does still frame the event description as intentionally conveyed, as important/relevant to the topic at hand, and as an "encapsulated message".

The second question raised is whether informational redundancy itself is sufficient to trigger such an inference. As mentioned previously, we start from the premise that rational speakers mention only that which cannot be automatically inferred by the comprehender. A charitable comprehender may be expected to expend considerable effort on rescuing the assumption of a cooperative or rational speaker (Davidson, 1974). If only activities under a certain threshold of habituality deserve mention, then comprehenders should conclude that the activity mentioned is relatively unusual, independently of any special emphasis on the utterance. In general, most types of inferences, if they occur, should occur as long as the semantic content of the utterance remain constant (cf. the "non-detachability" hypothesis).

On the other hand, pragmatic inferences must be calculable (Levinson, 2000), and utterances must be attended to closely enough in the

first place, before they may trigger any inferences (Sperber & Wilson, 2004). That is, particularly for non-generalized (context-sensitive) inferences, the context must offer sufficient support that the reader can infer the speaker's intent, or a plausible background state, with reasonable certainty. It is not clear, in our case, if the redundancy itself constitutes sufficient support. The degree of "intentionality" on the part of the speaker (also signaled in our stimuli by the exclamation mark) may also affect comprehenders' willingness and effort in guessing any implied meaning, as an utterance that may be a stray thought uttered without any specific intent may not be worth much effort to attempt to decipher (cf. the "form sensitivity" hypothesis). To test whether informational redundancy itself is sufficient for triggering the inference, or whether some amount of discourse or prosodic emphasis is necessary for its generation, we strip the event description of prosodic or discourse cues signaling speaker intentionality in Experiment 3.

3. Experiment 2: Implicit intent signaled by discourse markers

The second experiment tests whether the effect, of informationally redundant event descriptions being interpreted by readers as signaling activity *non-habituality*, is generalizable. To do so, we can replace the exclamation point with a non-prosodic discourse marker that signals speaker intentionality and utterance relevance (but crucially, not surprise). In this experiment, we frame the informationally redundant event description as an apparent recalling of information specifically intended to be mentioned to the comprehender, and implicitly relevant to the material just discussed: "*Oh yeah, and [he paid the cashier].*"

This discourse marker does not clearly signal surprise at the activity having been engaged in, nor does it explicitly support the intended inference otherwise – and in contrast to the exclamation mark in Exp.1, is a non-prosodic manipulation of the event description. We therefore consider it a good test of whether the effect generalizes beyond the specific context used in the first experiment.

3.1. Methods

3.1.1. Participants
700 eligible participants (787 total; median age bracket 26–35; 51.30% female) were recruited on Amazon Mechanical Turk. 87 participants were excluded from analysis (11.05%), following the same exclusion criteria as applied in Experiment 1.

3.1.2. Design
Experimental design was identical to experiment 1.

3.1.3. Materials
The same 24 stimuli were used as in Exp. 1. In this case, the critical utterance was prepended by "*Oh yeah, and…*", and the exclamation mark

was removed:

(2) OH YEAH, AND …

[1a] John often goes to the grocery store around the corner from his apartment$_{ordinary}$.	[1b] John is typically broke, and doesn't usually pay when he goes to the grocery store$_{wonky}$.
[2] 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.	
[3] Susan said to Peter:"John just came back from the grocery store.	
[4a] **Oh yeah, and** he *paid the cashier$_{habitual}$*."	[4b] **Oh yeah, and** he *got some apples$_{non-habitual}$*."

3.1.4. Procedure

The procedure was identical to that of Exp. 1.

3.1.5. Measures

The same response measures as in Exp. 1 were used to estimate *pre-utterance beliefs* and *post-utterance beliefs*.

3.2. Results

As in Experiment 1, to determine whether participants made inferences regarding activity habituality, we modeled *belief change* – the difference between *pre-utterance* and *post-utterance* beliefs. *Conventionally habitual* and conventionally *non-habitual* activities were again modeled separately. All factors were effect/sum coded.

3.2.1. Conventionally habitual activities

As we predicted, *pre-utterance belief* ratings for *ordinary context* activities showed that these activities are judged to be highly habitual (84.71). As in Experiment 1, *post-utterance beliefs* about the habituality of *ordinary context* activities were significantly lower (73.84), and *wonky* common ground estimates remained stable (47.45 *pre-utterance* to 47.47 *post-utterance*).

A linear mixed effects regression analysis, the results of which are summarized in Table 3, showed an interaction between context and belief measure ($\beta= -11.71$, $p<0.001$), which is driven by lowered activity habituality ratings when the readers see the utterance in a ordinary context ($\beta= -11.11$, $p<0.001$). All model specifications are as described in Exp. 1. A plot illustrating the interaction can be seen in Fig. 4, which shows a pattern of results that is remarkably quantitatively and qualitatively similar to that of Exp. 1. Exp. 1 and 2 are compared directly, and to Exp. 3, in Section 5.

These results support our prediction that readers perceive informationally redundant utterances as abnormal, and make pragmatic inferences (of activity *non-habituality*), regardless of whether implicit prosody or other markers conventionally associated with surprise are present.

3.2.2. Non-habitual activities

In contrast to Experiment 1, there was some increase in participants' ratings of *non-habitual* activities from *pre-utterance beliefs* (*ordinary*: 40.30 to 43.22; *wonky*: 37.74 to 43.05), see Fig. 5.

Table 3
Experiment 2: conventionally habitual (cashier-paying) activity analysis.

	β	SE(β)	t	p
Intercept	63.58	1.85	34.33	< 0.001
Common Ground: Ordinary	31.60	3.35	9.43	< 0.001
Belief: Post-utterance	−5.31	1.38	−3.83	< 0.001
Common Ground * Belief	−11.71	2.03	−5.76	< 0.001

A linear mixed effects regression analysis showed that estimates of activity habituality increase slightly when the utterance describing the *non-habitual* activity (see Table 4) is visible ($\beta=5.09$, $p<0.01$).

While not identical to the results of the first experiment (which showed a slight numerical increase in rating only), this is consistent with a peripheral prediction we made prior to running the experiments: simply mentioning a non-habitual, or non-redundant activity may increase the perception of its habituality, by providing some evidence that, e.g., *John* is at least an occasional *apple purchaser*. As the direction of this effect does not change our interpretation of the results, we leave it aside for future exploration.

3.3. Discussion

Together with Experiment 1, these results show that readers find informational redundancy abnormal at face value, and make pragmatic inferences to reconcile apparent informational redundancy with their expectations of utterance utility. This further disconfirms the "no inference" hypothesis, and indicates that the effect is generalizable, and not dependent on conventional indicators of activity non-habituality, such as implicit exclamatory intonation.

The results of Experiments 1 and 2, however, do not permit us to distinguish between the 2nd and 3rd hypotheses ("non-detachability" vs. "form sensitivity"), as they leave open the question of whether the *atypicality inference* effect is dependent on some degree of intentionality-signaling, or applies independently of discourse context. Experiment 2 provides some support for the "non-detachability" hypothesis, as the magnitude of the inference remains very stable, even as the form of intention or relevance signaling is substantially changed.

If the effect is dependent on some amount of relevance or intentionality signaling, this would support the "form sensitivity" hypothesis over the "non-detachability" hypothesis, by suggesting one of the following. Comprehenders may be relatively unwilling to expend substantial effort on decoding a plausible inference in the absence of evidence that doing so is worth it, and that the utterance has some amount of import. Similarly, they may stop short in their efforts, on the assumption that it is more likely that speakers would occasionally violate this particular conversational maxim, than that they would provide insufficient evidence that the utterance communicates something of note. Finally, they may simply be generally tolerant of informational redundancy, unless context suggests that the redundancy has a "point." Experiment 3 presents the same task and materials to participants, but removes the exclamation mark or discourse marker that signals relevance and speaker intent.

4. Experiment 3: Removing evidence of speaker intent

To investigate whether explicitly signaling speaker intent has an influence on the strength of the *atypicality inference* effect, we designed a third experiment which differs only in the absence of the exclamation mark or a discourse marker, and contains hence no special signals for the relevance/informativity of the activity description. Our prediction is that while the effect may be attenuated somewhat, comprehenders should nevertheless make a measurable attempt to compensate for a violation in expected informational utility (i.e., while there may be some degree of "form sensitivity," the inference should nevertheless arise).

4.1. Methods

4.1.1. Participants

700 eligible participants (759 total; median age bracket 26–35; 51.60% female) were recruited on Amazon Mechanical Turk. 59 participants were excluded from analysis (7.77%), following the same exclusion criteria as applied as in previous experiments.
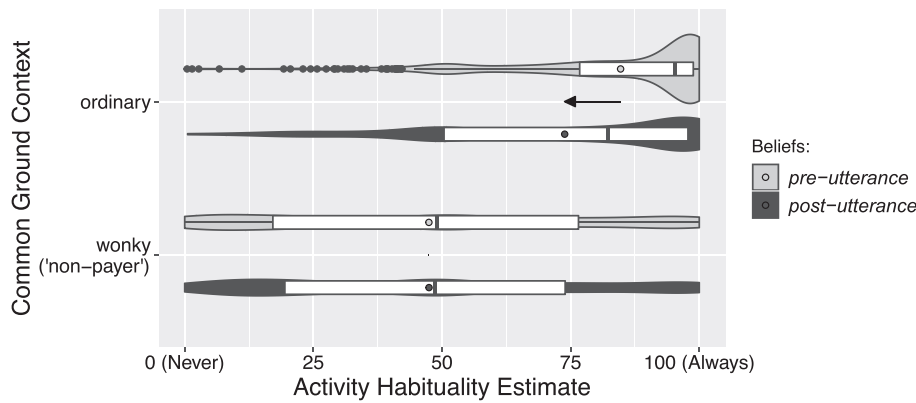
Fig. 4. Experiment 2: *conventionally habitual* (*cashier-paying*) activity analysis.
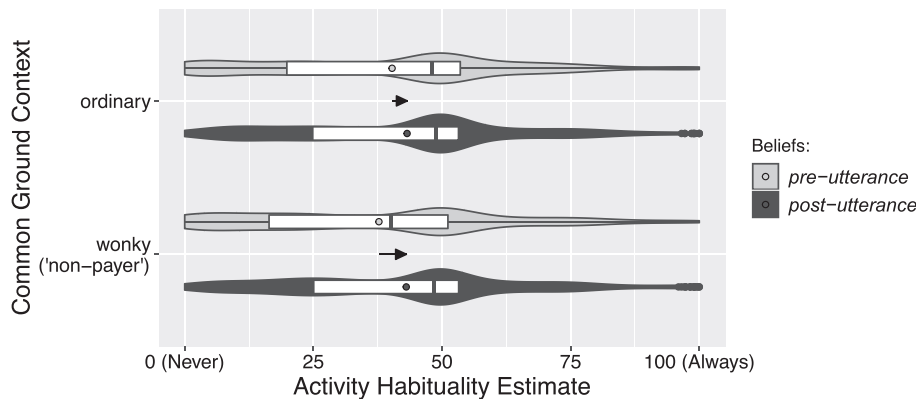

Fig. 5. Experiment 2: *non-habitual* (*apple-buying*) activity analysis.

Table 4
Experiment 2: non-habitual (apple-buying) activity analysis.

	β	SE(β)	t	p
Intercept	40.99	1.85	22.14	< 0.001
Common Ground: Ordinary	0.95	1.83	0.52	0.6
Belief: Post-utterance	5.09	1.78	2.86	< 0.01
Common Ground * Belief	−1.22	1.55	−0.79	0.4

4.1.2. Design

The design was the same as for experiments 1 and 2.

4.1.3. Materials

The same 24 stimuli were used as in the previous experiments. The only alteration from Experiment 1 was the substitution of the exclamation point with a period.

4.1.4. Procedure

The procedure was identical to that of previous experiments.

4.1.5. Measures

The same response measures as in the previous experiments were used to estimate *pre-utterance beliefs* and *post-utterance beliefs*.

4.2. Results

As in previous experiments, we modeled the difference between *pre-utterance* and *post-utterance* beliefs. *Conventionally habitual* and *non-habitual* activities were modeled separately. All factors were effect/sum coded.

4.2.1. Conventionally habitual activities

As in the previous experiments, *pre-utterance belief* ratings showed *ordinary context* activities to be highly habitual (85.59), and *wonky context* activities to be less habitual (49.50), see Fig. 6. Consistent with our predictions, *post-utterance beliefs* are significantly lower in the *ordinary context condition* (80.30), but less so than in the previous two experiments. Exp. 3 is compared directly to Exp. 1 and 2 in Section 5.

A linear mixed effects regression analysis, the results of which are summarized in Table 5, showed a statistically significant interaction between context and belief measure ($\beta= -5.40$, $p<0.01$), which is driven by lowered activity habituality ratings when the readers see the utterance in an ordinary context ($\beta= -4.87$, $p<0.001$). All model specifications are as described in Exp. 1 and 2.

These results indicate that, consistent with our predictions and the results of Exp. 1 and 2, when an easily inferable activity is overtly mentioned in a *ordinary* common ground context, comprehenders do infer some degree of activity non-habituality, even without implicit prosody or discourse markers putting additional emphasis on the utterance.

4.2.2. Non-habitual activities

In contrast to Experiment 1 and similar to Experiment 2, there was some increase in participants' ratings of *non-habitual* activities from *pre-utterance* to *post-utterance* beliefs (*ordinary*: 41.08 to 46.46; *wonky*: 37.61 to 44.42), see Fig. 7.

A linear mixed effects regression analysis showed that estimates of activity habituality do not vary with changes in the common ground context (or common ground *wonkiness*), but do increase slightly when the utterance describing the *non-habitual* activity (see Table 6) is visible ($\beta=6.88$, $p<0.001$). As in the case of Exp. 2, we suspect that explicitly mentioning a relatively unusual activity leads participants to believe
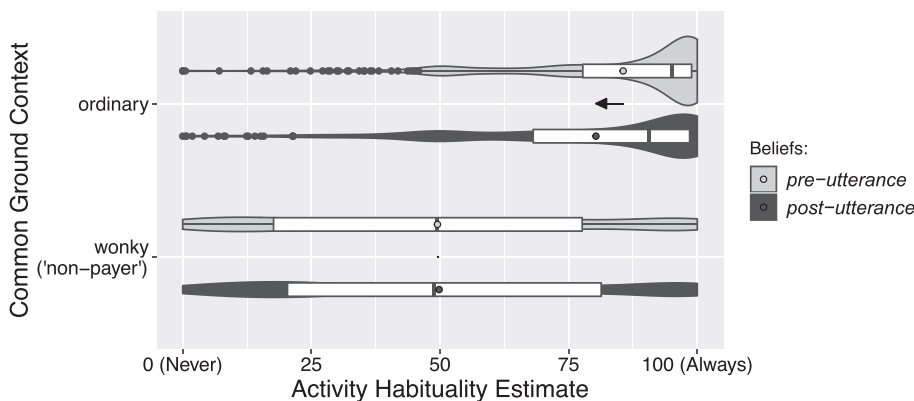
Fig. 6. Experiment 3: *conventionally habitual* (*cashier-paying*) activity analysis.

Table 5
Experiment 3: conventionally habitual (cashier-paying) activity analysis.

	β	SE(β)	t	p
Intercept	66.38	1.88	35.40	< 0.001
Common Ground: Ordinary	33.21	3.40	9.77	< 0.001
Belief: Post-utterance	−2.20	0.93	−2.36	< 0.05
Common Ground * Belief	−5.40	1.75	−3.10	< 0.01

that activity to be slightly more habitual than they would otherwise assume.

4.3. Discussion

In addition to the results of the first two experiments, these results suggest that, when informationally redundant utterances are presented without a signal of speaker intent and utterance relevance, comprehenders are relatively unlikely to draw atypicality inferences. This is consistent with the "form sensitivity" hypothesis described in Section 1.4, and the premise that such inferences are dependent on the degree to which the utterances are perceived as intentional.

Further, the fact that there is an effect across all three experiments means that there is also some support for the non-detachability hypothesis. Note though that the strong form of the "non-detachability" hypothesis is not supported, as the effect is smaller in the absence of the explicit intention cues. We conclude that both form and content matter.

5. Cross-experiment analysis

In this section, we directly compare the effect sizes of the atypicality inference effects across all three experiments. We run a $3 \times 2 \times 2$ linear

mixed effects regression analysis of *conventionally habitual* activities. We modeled *belief change* (*pre-utterance* vs. *post-utterance* beliefs), as a function of common ground (*ordinary* vs. *wonky*), as well as the between-subject discourse marker manipulation ("*!*" vs. "*Oh yeah, and*" vs. "*.*"). The first two factors were effect/sum coded. We used Helmert coding for the 3-level experiment factor, as this allowed us to make the comparisons of theoretical interest: Exp. 1 vs. Exp. 2 ("*!*" vs. "*Oh yeah, and*"), and then Exp. 3 vs. Exp. 1 and 2 grouped together ("*.*" vs. the relevance markers).

We used the maximal converging model, with by-subject random intercepts and slopes for common ground context (*ordinary / wonky*) and belief measure (*pre-utterance / post-utterance*), by-item random intercepts and slopes for both factors and their interaction, and a by-item random slope for experiment. By-subject random slopes for the interaction were not included in the model due to lack of within-subject repeated measures. The random slope for the full (by-item) experiment by common ground by belief measure interaction was not included due to non-convergence.

The regression analysis (see Table 7) showed a significant three-way interaction between relevance marker presence, common ground context, and belief measure: there was a significantly smaller *atypicality inference* effect in Experiment 3 than in Experiments 1 and 2 (β=5.78,

Table 6
Experiment 3: non-habitual (apple-buying) activity analysis.

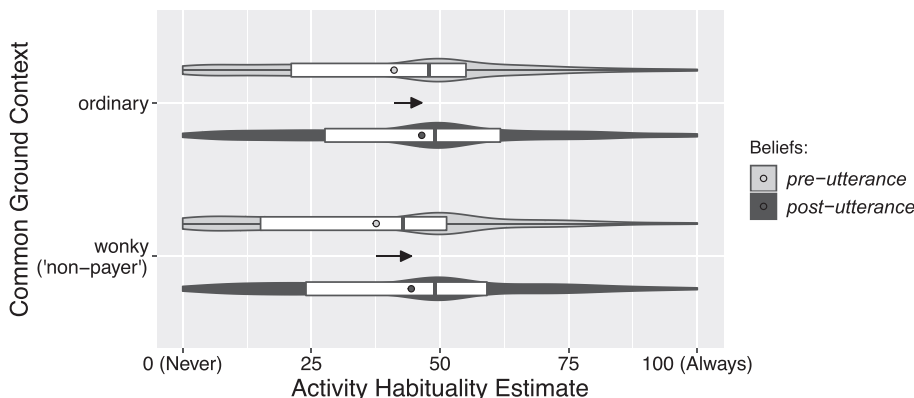|	β	SE(β)	t	p
Intercept	42.12	2.12	19.84	< 0.001
Common Ground: Ordinary	2.29	2.41	0.95	0.4
Belief: Post-utterance	6.88	1.77	3.88	< 0.001
Common Ground * Belief	−1.39	1.72	−0.81	0.4

Fig. 7. Experiment 3: *non-habitual* (*apple-buying*) activity analysis.

Table 7
Experiments 1–3: conventionally habitual (cashier-paying) activity analysis.

	β	SE (β)	t	p
Intercept	64.34	1.78	36.14	**< 0.001**
'!' vs. 'Oh yeah…'	0.49	0.86	0.57	0.6
'.' vs. Relevance Markers	3.08	0.81	3.81	**< 0.001**
Common Ground: Ordinary	32.40	3.22	10.05	**< 0.001**
Belief: Post-utterance	−5.07	1.04	−4.86	**< 0.001**
'!' vs. 'Oh yeah' * Common Ground	−0.64	1.43	−0.45	0.7
'.' vs. Relevance Markers * Common Ground	1.25	1.24	1.01	0.3
'!' vs. 'Oh yeah' * Belief	2.50	1.30	1.92	0.1
'.' vs. Relevance Markers * Belief	4.52	1.13	4.01	**< 0.001**
Common Ground * Belief	−9.33	1.11	−8.41	**< 0.001**
'!' vs. 'Oh yeah' * CG * Belief	−0.60	2.35	−0.26	0.8
'.' vs. Relevance Markers * CG * Belief	5.78	2.03	2.84	**< 0.01**

$p{<}0.01$), and no significant difference between Experiments 1 and 2 ($\beta = -0.60$, $p{=}0.80$).

As predicted, the atypicality inference effect holds regardless of which relevance marker is used, and in fact there is no statistically significant difference between the two markers. Further, the effect size of the common ground by belief measure interaction is significantly smaller in the absence of the markers; in other words, participants are significantly less likely to make a *atypicality* inference in the absence of an exclamation mark or a discourse marker signaling relevance or intentionality.[12]

This result clearly favors the "form sensitivity" hypothesis described in Section 1.4 over a strong version of the "non-detachability" hypothesis (which might predict an effect of the same magnitude for all experiments). We conclude that in the absence of a clear signal of utterance relevance or speaker intentionality, comprehenders are either less likely to attempt to resolve the violation, resolve it in a manner that is not reflected in our response measures, or do not detect the violation the first place. The first possibility is supported by observations that comprehenders approach speaker utterances *charitably*, and may expend significant effort on interpreting them in a manner that is consistent with the speaker making cooperative conversational choices (Davidson, 1974). However, it is also possible that comprehenders are less "charitable" in general when presented with oddly phrased psycholinguistic stimuli in an artificial setting – as well as less motivated on expending cognitive effort on calculating a non-obvious inference in a non-interactive environment, on the basis of an utterance that their attention is not otherwise drawn to.

Less charitable comprehenders, who may detect the redundancy but fail to in some way resolve it, may assume that the speaker is odd or not a particularly cooperative speaker, or perhaps that they are having production difficulties. Another possibility is that they assume the speaker is in the process of planning a more informative utterance (where, for example, the description might serve as a temporal/causal anchor; see Example 8). Determining which strategies comprehenders do in fact resort to, and in which contexts, is left to future work.

6. Is the effect of habituality on pragmatic inferences gradient?

Finally, we would like to analyse whether the effect of redundancy on atypicality inferences is a gradient one, or whether there is any evidence of a categorical difference between highly predictable and less predictable events. From a theoretical point of view, there has been a distinction between overinformtive uttances and utterances which are not overinformative. For instance, saying "large red cup" when "large

cup" would be sufficient to pick out the correct target. In the cases of our stimuli, the distinction between an overinformative event and an informative one is more gradual: some of the target events in our stimuli are estimated to be highly likely ($p > 0.9$) to usually happen, while others are estimated to have a probability that is just slightly higher than the most probable ones of our non-habitual events. We therefore think that it would be interesting to see whether the observed effects are graded, or whether there is a clearer decision boundary between over-informative utterances and non-overinformative ones. A recent paper by Degen, Hawkins, Graf, Kreiss, & Goodman, 2020 similarly proposes that under their continuous semantics RSA account, the distinction between even the traditional categorial notion of overinformativity should be replaced by a graded notion, where the degree of overinformativity depends on the communicative benefit of mentioning an attribute (which is conceptualized in their work as world-knowledge related noise). In this section, we therefore analyse the data from experiments 1–3 on an item-by-item basis.

Fig. 8 plots the measured average activity habituality, with and without seeing the target utterance, for each item in each condition, for all three experiments. The diagonal dashed line demonstrates what the "no inference" hypothesis would predict: i.e., no effect of the utterance on belief change (*pre-utterance* ratings mapping straightforwardly onto *post-utterance* ratings). Points found above the line indicate that for those items, participants were more certain, for example, that *John usually buys apples* when the story mentioned that "he got some apples." Points below the line indicate *an atypicality inference*: e.g., mentioning that "he paid the cashier" causes people to believe that *John does not usually pay the cashier*.

In Experiment 1 (exclamation mark), we see that for *ordinary* common grounds, and *conventionally-habitual* activities (e.g., *paying the cashier* given an ordinary common ground), most data points fall below the line, indicating an atypicality inference. Interestingly, we also see a gradient "trend" towards *non-habituality* in the other three (non-redundant) conditions: items that are similar to *ordinary habitual* items, in terms of pre-utterance habituality estimates, are more likely to trigger atypicality inferences. In contrast, items with low pre-utterance habituality estimates show the opposite effect: i.e., if it is mentioned that an individual engaged in a particularly non-habitual activity, it leads comprehenders to believe that the individual is more likely to engage in that activity habitually. The same observations also hold for Experiment 2.

In Experiment 3 (period), we again see a graded effect of *pre-utterance* beliefs regarding activity habituality on the likelihood of a *atypicality inference*, but this time the regression line is shifted upwards (Exp. 1: $\beta{=}0.64$; Exp. 2: $\beta{=}0.64$; Exp. 3: $\beta{=}0.76$). We still see, however, that there is a gradual difference between highly expected vs. relatively expected events, in terms of likelihood of an atypicality inference occurring.

Taken together, we can see in these figures that the exclamation mark and the "*oh yeah…*" discourse marker, as signals of speaker effort and intentionality, make it more likely that atypicality inferences will arise for *ordinary* common ground, *habitual* activity activity mentions. Furthermore, we can see that the effect of pre-utterance beliefs on atypicality inferences is clearly gradient rather than binary: relatively more habitual activities, in all conditions, generally elicit larger *atypicality* inferences.

7. General discussion and conclusion

Taken together, this series of experiments shows that comprehenders react to informationally redundant utterances by shifting their beliefs about the common ground, such that the utterances are more "informative" in context, thus increasing their utility. This occurs even though informational redundancy, or overinformativity, in itself has no obvious negative impact on basic message comprehension. This is consistent with theoretical accounts of what constitutes "cooperative"

---

[12] A similar cross-experiment analysis of the conventionally *non-habitual* activities can be found in the supplementary materials: https://osf.io/8fz4m/?view_only=ff5859d3f33b485d95254395f95a52dc
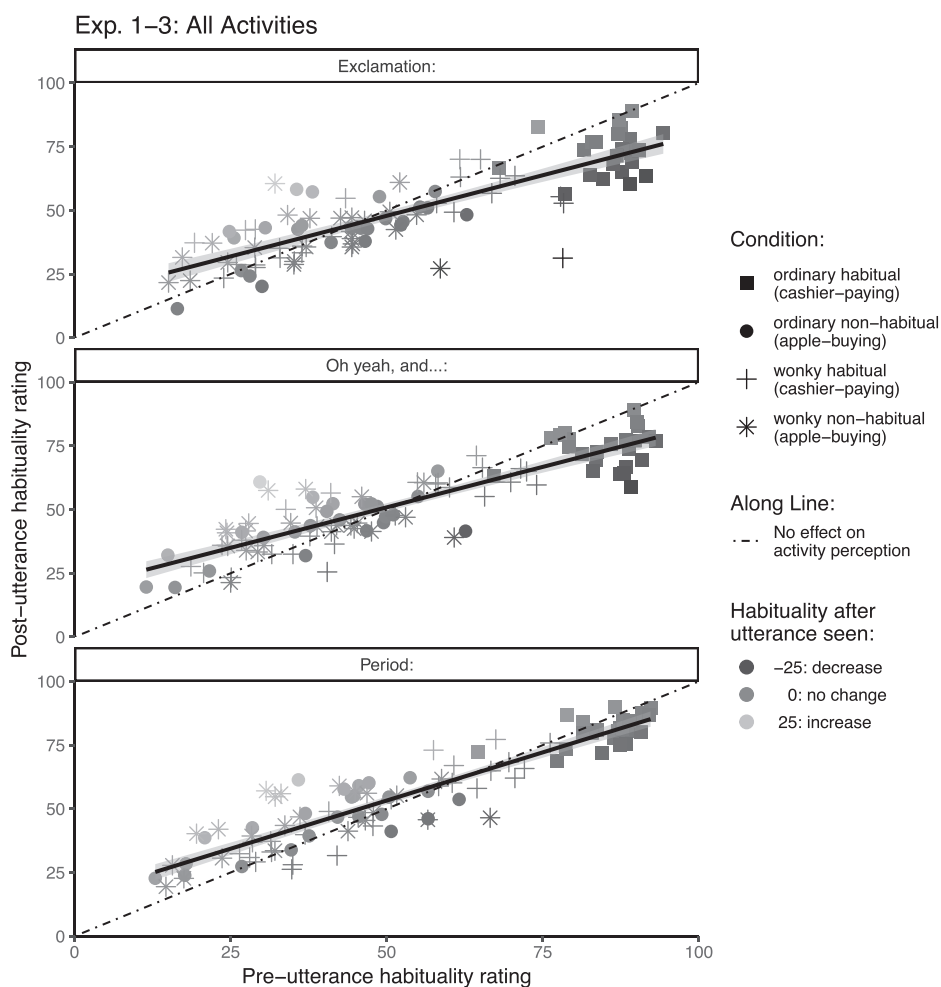
Fig. 8. These plots show by-item belief change for all conditions of our three experiments. We observe a highly gradient effect of activity typicality on the likelihood to draw an atypicality inference.

communicative behavior (Grice, 1975), as well as of comprehenders' attempts to resolve speaker behavior that at face value does not appear particularly rational. However, as the third experiment shows, the effect is significantly modulated by how the utterance is framed in the discourse, supporting the hypothesis that inference strength is sensitive to utterance form. Overall, we provide robust evidence that informational redundancy is perceived as anomalous, and that comprehenders alter their situation models to accommodate it, particularly when there's evidence that there was specific intent behind the utterance.

While previous work on informationally redundant utterances focused on redundant modifications such as 'grey elephant' or "*the **long** fork*" in the absence of any other fork, our work here contributes to the discussion by testing substantially longer and more costly redundant utterances, and thereby contributes an interesting new datapoint regarding the conditions under which overinformative utterances can give rise to pragmatic inferences.

Another area of contribution is that we illustrate a case in which comprehenders are willing to *revise the assumed common ground* of the discourse, in order to accommodate a perceived violation in the informational utility of an utterance. Unlike shifting assumptions about intended utterance meaning, this is a strategy that has not received much attention to date, with the notable exception of Degen et al. (2015). The shifting of common ground assumptions appears to be an important, and surprisingly understudied strategy for interpreting utterances that, at face value, violate conversational norms. Neglecting it as a possibility risks leading to misinterpretation of online effects and under-detection of pragmatic inferences in experimental work.

We show that semantically "vacuous" utterance features (those that do not alter the propositional content of an utterance), in the form of implicit prosody or discourse markers, significantly influence the extent to which comprehenders draw an inference predicted by pragmatic theories of rational speaker behavior. Aside from the case of contrastive prosody (Bergen & Goodman, 2015; Kurumada, Brown, & Tanenhaus, 2012; Ward & Hirschberg, 1985), this has not to date been systematically investigated in formal or experimental literature, and most likely also extends to other pragmatic phenomena. In our case, we argue that comprehenders are weighing and evaluating multiple cues regarding how likely it is that a speaker intended to communicate a particular meaning, or that the common ground or background state is substantially different from what was initially assumed, and should be revised.

We would also like to discuss why we call the pragmatic inferences observed in these experiments simply inferences and not implicatures. A core aspect of an implicature is that the message behind it must have been intended by the speaker to be understood by the addressee. In the stimuli used in the present experiments, we do not make claims about whether the speaker actually intended the inference or not. Our experimental participants are in a position of a third-party comprehender who is not privy to the background knowledge of the speaker. And the speaker who utters the critical utterance in our stories is not addressing the participants. This lack of intentionality thus does not qualify the inferences observed here as implicatures. We believe that it is nevertheless important to be able to model the change between pre-utterance and post-utterance beliefs about the common ground, given that this change can have a marked effect on which inferences are drawn by

comprehenders (see also Degen et al., 2015). An exclusive focus on intended meanings, rather than changes in background assumptions, may lead to erroneous conclusions that comprehenders are drawing no pragmatic inferences from an given utterance.

There are several avenues for further research. First, the range of inferences that comprehenders might draw from informationally redundant utterances may extend well beyond what we tested in this series of experiments. For instance, in the absence of a possible pragmatically felicitous interpretation, as the one suggested by our response measure, comprehenders may simply assume that a speaker is being uncooperative, having some production difficulty, or has unconventional speaking patterns (cf. Grodner and Sedivy, 2011; Pogue et al., 2016). There is also the possibility that informationally redundant event descriptions, especially as seen in Experiment 3, are initially interpreted as likely, and possibly aborted, temporal or causal anchors for more "interesting" information. For example, in the context of a *grocery trip*, an "informationally redundant" description such as *John paid the cashier*, when followed by *with euros instead of dollars*, would likely not be considered anomalous. In this case, the description would not be redundant in its broader context, as it is part of a more extended description that overall contributes previously unknown, or not easily inferable information. These hypotheses might be investigated using rating studies, sentence or passage completion studies, or more naturalistic tasks where participants' behavior provides a clue as to their interpretation of these utterances. For instance, one could ask them about what they think the speaker intended to communicate with the redundant utterance. Alternatively, one could try to elicit explanations of *why* they put the slider bar at a specific position. From those explanations, we would then be able to understand whether they actually made an inference about the typicality of the event, or whether they accommodated the redundant utterance in a different way.

Secondly, it would be interesting to explore whether pragmatic inferences raised by informationally redundant utterances are cognitively effortful. There is a long-standing debate not only as to whether redundant utterances give rise to pragmatic inferences, but also as to whether these inferences are cognitively effortful. Most pragmatic accounts would predict that this should be the case for atypicality inferences, given that they are a type of particularized implicature (Levinson, 2000), and given that the target inference (non-habituality of John paying the cashier when shopping) is not directly contextually primed or supported (Degen & Tanenhaus, 2016; Wilson & Sperber, 2002). Given that studies on the effortfulness of pragmatic inferences have to date mostly focused on scalar implicatures (see e.g., Dieussaert, Verkerk, Gillard, & Schaeken, 2011; Grodner, Klein, Carbary, & Tanenhaus, 2010; Bott & Noveck, 2004; Bott, Bailey, & Grodner, 2012; Huang & Snedeker, 2009; Marty, Chemla, & Spector, 2013), testing the effortfulness of atypicality inferences could add a very interesting data point to the theoretically important question of whether a measurable amount of cognitive effort can be detected for particularized inferences, and hence potentially shed light on how such inferences are derived.

Finally, it should be explored how the effects reported in our studies (the overall atypicality inference effect, as well as the differences in effect size between the full-stop condition and the exclamation mark or discourse-marked but meaning-equivalent utterances) can be accounted for by models of pragmatics. We here briefly sketch how a Rational Speech Act (RSA) model could be configured to capture the effects demonstrated in this article; for more details, please refer to Kravtchenko and Demberg (2022). Firstly, we note that the RSA base model would need to be extended with a joint reasoning component regarding the habituality of activities, in order to model both changes in beliefs about the world, and changes in beliefs about the event mentioned in the text. Similar mechanisms have previously proposed by Degen et al. (2015) and Goodman and Frank (2016), this mechanism could hence be adapted to atypicality inferences about habituality. Empirical priors for the likelihood of activity habituality (corresponding to the pre-utterance beliefs estimated in our experiments) could be fed into the model.

A second crucial observation is that a standard RSA model principally cannot predict inferences of different strengths for the different utterance prominence conditions. The failure of standard RSA models to derive pragmatic inferences of different strengths, given semantically meaning-equivalent utterances, is directly analogous to their failure to derive M-implicatures or inferences due to prosodic stress, as detailed and mathematically proven in Bergen, Levy, and Goodman (2016). In order to capture effects of utterance prominence, it is necessary to assign some attentional or memory-related benefit to the more costly redundant utterance (here, the one with the exclamation mark or the discourse marker), to be already active at the literal listener level. Empirically, there is evidence that readers often cannot recall whether elements in a stereotyped activity sequence were explicitly mentioned, or not (Bower et al., 1979), and that informational redundancy, even at the multi-word level, in part serves the purpose of ensuring that listeners attend to and accurately recall relevant information (Baker et al., 2008; Walker, 1993). The noisy-channel RSA model proposed by Bergen and Goodman (2015), with minimal modification, could capture this intuition, although in our case the "noise" would relate to whether an utterance is attended to and stored in memory, rather than whether it is misheard.

CRediT authorship contribution statement

Ekaterina Kravtchenko: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Vera Demberg:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

References

Arnold, B. F., Hogan, D. R., Colford, J. M., & Hubbard, A. E. (2011). Simulation methods to estimate design power: An overview for applied research. *BMC Medical Research Methodology, 11*(94).

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics, 43*(1), 361–374.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech, 47*, 31–56.

Baker, R., Gill, A., & Cassell, J. (2008). *Reactive redundancy and listener comprehension in direction-giving. Columbus, Ohio*.

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

Bergen, L., & Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. *Topics in Cognitive Science, 7*(2), 336–350. https://doi.org/10.1111/tops.12144

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics, 9*.

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

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language, 51*(3), 437–457.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology, 11*, 177–220.

Cohen, P. R. (1978). *On knowing what to say: Planning speech acts*. Doctoral Dissertation. University of Toronto.

Davidson, D. (1974). Belief and the basis of meaning. *Synthese, 27*, 309–323.

Davies, C., & Katsos, N. (2010). Over-informative children: Production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua, 120*, 1956–1972.

Davies, C., & Katsos, N. (2013). Are speakers and listeners "only moderately Gricean"? An empirical response to Engelhardt et al. (2006). *Journal of Pragmatics, 49*(1), 78–106.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review, 127*(4), 591.

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science, 40*(1), 172–201.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci2015)* (pp. 548–553).

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

Engelhardt, P. E., Bailey, K. G. D., & Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language, 54*, 554–573.

Fillmore, C. J. (2006). Frame semantics. In *Encyclopedia of language and linguistics* (pp. 613–620). https://doi.org/10.1016/B0-08-044854-2/00424-7

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*(6084) (998–998).

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*(11), 818–829.

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Speech acts: Vol. 3. Syntax and semantics* (pp. 41–58). New York: Academic Press.

Grodner, D., & Sedivy, J. C. (2011). In N. Pearlmutter, & E. Gibson (Eds.), *The processing and acquisition of reference* (pp. 239–272). Cambridge, MA: MIT Press.

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

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form and use in context* (pp. 11–42). Washington: Georgetown University Press.

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

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61*(1), 23–62.

Kravtchenko and Demberg. (2022). *Modeling atypicality inferences in pragmatic reasoning. Ekaterina Kravtchenko and Vera Demberg.* Science Society (CogSci 2022).

Kravtchenko, E., & Demberg, V. (2015). Semantically underinformative utterances trigger pragmatic inferences. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci2015)* (pp. 1207–1212).

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2012). Pragmatic interpretation of contrastive prosody: It looks like speech adaptation. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci2012)* (pp. 647–652).

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26.

Levinson, S. C. (2000). *Presumptive meanings - the theory of generalized conversational implicature.* The MIT Press.

Long, M., Rohde, H., & Rubio-Fernandez, P. (2020). The pressure to communicate efficiently continues to shape language use later in life. *Scientific Reports, 10*(1), 1–13.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition, 126*(2), 313–318.

Mangold, R., & Pobel, R. (1988). Informativeness and instrumentality in referential communication. *Journal of Language and Social Psychology, 7*(3–4), 181–191.

Marty, P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua, 13*(3), 152–163.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science, 13*, 329–336.

Paraboni, I., & van Deemter, K. (2014). Reference and the facilitation of search in spatial domains. *Language, Cognition and Neuroscience, 29*(8), 1002–1017.

Paraboni, I., Van Deemter, K., & Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational Linguistics, 33*(2), 229–254.

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Frontiers in Psychology, 6*(2035).

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).

Rehrig, G., Cullimore, R. A., Henderson, J. M., & Ferreira, F. (2021). When more is more: Redundant modifiers can facilitate visual search. *Cognitive Research: Principles and Implications, 6*(1), 1–20.

Rett, J. (2011). Exclamatives, degrees and speech acts. *Linguistics and Philosophy, 34*(5), 411–442.

Rohde, H., Futrell, R., & Lucas, C. G. (2021). What's new? A comprehension bias in favor of informativity. *Cognition, 209*, Article 104491.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology, 7*, 153.

Ryzhova, M., & Demberg, V. (2020). Processing particularized pragmatic inferences under load. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci 2020)*.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research, 32*(1), 3–23.

Sonnenschein, S., & Whitehurst, G. J. (1982). The effects of redundant communications on the behavior of listeners: Does a picture need a thousand words? *Journal of Psycholinguistic Research, 11*(2), 115–125.

Sperber, D., & Wilson, D. (2004). *Relevance theory. Handbook of Pragmatics* (pp. 607–632). Blackwell: Oxford.

Tourtouri, E. N., Delogu, F., Sikos, L., & Crocker, M. W. (2019). Rational over-specification in visually-situated comprehension and production. *Journal of Cultural Cognitive Science, 3*(2), 175–202.

Walker, M. A. (1993). Informational redundancy and resource bounds in dialogue. In *Stanford university; Doctoral Dissertation*. Philadelphia, PA: University of Pennsylvania.

Ward, G., & Hirschberg, J. (1985). Implicating uncertainty: The pragmatics of fall-rise intonation. *Language, 61*, 747–776.

Wilson, D., & Sperber, D. (2002). Relevance theory. In L. R. Horn, & G. Ward (Eds.), *Vol. 1. The handbook of pragmatics* (pp. 606–632). Oxford, UK: Blackwell Publishing. https://doi.org/10.1002/9780470756959.ch27

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 386–397.