Informationally Redundant Utterances Alter Prior Beliefs about Event Typicality

Abstract Most theories of pragmatics and language processing predict that speakers avoid excessive informational redundancy. Informationally redundant utterances are, however, quite common in natural dialog. From a comprehension standpoint, it remains unclear how listeners 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 listener expectations. In a series of 3 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 typical, such as paying the cashier when shopping, is reinterpreted as atypical). Further, we show that the degree to which such inferences are triggered depends on the framing of the utterance. In the absence of prosodic or discourse markers 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 listeners revising the conversational common ground, in an effort to accommodate unexpected dips in informational utility.

Keywords: Psycholinguistics; redundancy; context-dependent implicatures; accommodation; overinformativeness; prosody.

Word count (excluding numbers): 17,765

1 Introduction

Pragmatic theories and theories of language processing typically include constraints against utterances which add no new information to the discourse, or are informationally redundant (Aylett & Turk 2004, Cohen 1978, Grice 1975, Jaeger 2010). 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. Examples include vowel shortening (Aylett & Turk 2004), use of shorter word variants (Mahowald et al. 2013), or omission of optional complementizers
(Jaeger 2010). At the informational or conceptual level, redundancy refers to the explicit mention of information that could be automatically inferred by the listener using world knowledge or common ground information, or that is already entailed or strongly implied by the preceding discourse.

In contrast to redundancy at the form level, constraints against overinformativeness, or redundancy, at the informational level have always been debated (Grice 1975). There is ample evidence that these are routinely violated, and frequently tolerated (Baker et al. 2008, Engelhardt et al. 2006, Nadig & Sedivy 2002, Walker 1993). In this paper, we explore the question of whether such constraints do in fact exist, when they come into play, and how listeners react to and interpret apparent violations. Specifically, we look at cases where the redundancy is at the level of world knowledge. Utterances such as (1) are redundant at face value, in that they overtly state that ‘John’ paid the cashier, which conventionally can be inferred simply on the basis of him having gone shopping.

(1) “John went grocery shopping. He paid the cashier.”

Once the first sentence is uttered, the probability of a world state where the paying action has occurred is extremely high. 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. However, in this paper we show that listeners accommodate these utterances by changing their prior beliefs about the world state.

Utterances like the one shown in (1) are redundant on the basis of context and world knowledge. As context and background knowledge are fairly unsystematic and can be difficult to control for, we here use script, or schema knowledge as a proxy for 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 (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 part of its conventional meaning, and explicitly mentioning their occurrence would therefore be redundant.  

---

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 not often mentioned overtly (Bower et al. 1979, Regneri et al. 2010).
Utterance (1) introduces a well-known script or event sequence (grocery shopping), and the informationally redundant utterance (he paid the cashier!) consists of describing a highly predictable sub-event from this script. In this example, the event described in the second utterance is already conventionally implied to have occurred, by the preceding invocation of the grocery shopping script, as the assumption shared by most speakers and listeners is that people overwhelmingly pay cashiers when they go grocery shopping. Mentioning it explicitly, therefore, is informationally redundant, and if the utterance is to conform with listener expectations about speaker informativity, then the listener is likely to wonder what other information the speaker may have been trying to convey.

1.1 Informational Redundancy

While mainstream pragmatic theories do address cases where the information contained in an utterance may be redundant, they largely leave open the questions of whether listeners do, in fact, perceive redundancy as pragmatically infelicitous, as well as how they treat such utterances. Most theories are based on the idea that listeners expect speakers to behave rationally——namely, by communicating in a way that is consistent with communicative goals. However, as Grice (1975) notes, it’s unclear whether excessive redundancy comes into any real conflict with the goal of successful (truthful, sufficiently informative, relevant, etc.) communication—although listeners 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). Under this definition, it is not straightforwardly clear whether overinformativeness constitutes non-rational speaker behavior, and specifically whether this part of the Quantity maxim holds: “do not make your contribution more informative than is required.”

It is, however, possible that listeners perceive excessive information as, at minimum, non-relevant to the discourse, and therefore as violating the maxim of Relation (also speculated by Grice). The question, then, is whether listeners make any particular note of redundancy, find it odd or infelicitous, or attempt to accommodate it. If listeners 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, listeners ought to interpret conceptually redundant utterances as either an attempt to convey some non-literals (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 listeners in fact react to overinformativeness, and, by extension, the extent to which they expect speakers to be concise, has to date only been investigated within the relatively narrow scope of nominal modification, where the evidence, discussed further in Section 2, has largely been equivocal.

Ultimately, the question of how listeners react to overinformativeness is relevant to a more general theory of human communication, and must be answered to determine the extent to which: a) listeners and/or speakers behave in a ‘Gricean’ manner; b) under which conditions they do so, and which deviations from idealized communicative norms they are more likely to violate; and c) how and whether listeners resolve apparent violations, and which assumptions about implied meaning or common ground they are willing to make, in order to do so. If listeners do not appear to make much of overinformativeness (whether in terms of inferences made or possible violations perceived), and there is no evidence that speakers use overinformative utterances to convey specific non-literal meanings, then it’s questionable to what degree overinformativeness violates communicative norms or Gricean maxims, in the first place.

1.2 How might speakers react to informational redundancy?

In this section, we speculate in more detail how comprehenders might react to informational redundancy, or overinformativeness. We distinguish three theoretical possibilities. Specifically, we will consider what might happen when a comprehender encounters one of our experimental utterances (normally embedded within a larger context):

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

First, 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 be mediated entirely by how redundant the event being described is, independently of any other factors.
The second possibility is that listeners do not find informational redundancy particularly marked, as it typically does not interfere with comprehension—or, at most, find redundant utterances slightly odd or suboptimal, as has been found in some studies (Davies & Katsos 2010). It’s both likely that listeners do not expect speakers to exhibit optimally rational behavior at all times, and that communicative norms which have little or no effect on the listener’s ability to understand the basic meaning of an utterance, to the extent that they exist, are rather weak. Similarly, comprehenders may note redundancy in speech, but not draw any inferences regarding the intended meaning. They might instead ascribe the redundancy to some kind of speaker error: perhaps the speaker is stalling for something else to say, having production difficulty, or is simply not communicating optimally in that particular instance. In the case of our utterance, in this scenario, we would 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. As in the previous scenario, comprehenders’ reactions should be mediated only by how redundant the event described is—i.e., other factors should make no difference in how redundancy is interpreted.

There is a third possibility, suggested by the argument that implicatures must be calculable, or sufficiently supported by the context, such that a listener would not encounter undue difficulty in inferring the intended meaning (Levinson 2000). Ability or willingness to resolve apparent violations of conversational norms may be a function of the amount of effort listeners must expend on doing so (Sperber & Wilson 1995), as well as the amount of evidence that listeners have of the speaker having intended to say something ‘informative’ in the first place. For example, if a speaker is habitually overinformative, the listener has additional reason to suppose that a redundant utterance (‘John paid the cashier.’) carries no extra meaning. In contrast, if the speaker is in general reasonably concise, and further, if their speech contains other cues that the redundant utterance is describing a surprising or noteworthy event (e.g., using exclamatory prosody: ‘John paid the cashier!’), the listener has more reason to suspect that the otherwise apparently predictable event is, in fact, noteworthy and unexpected. In this scenario, whether listeners draw an inference from an informationally redundant utterance, or not, would be determined not just by the redundancy of the utterance, but also by the amount of support, from context and/or experience, for any given inference. This is consistent with the idea that communicative norms can be both exploited (to communicate something ‘extra’ through an apparent violation); and flouted, or straightforwardly violated. Presumably, listeners must rely on context and prior experience to determine which is which.
In this paper, we present three experiments, run concurrently on the same population, which test whether informationally redundant event descriptions lead comprehenders to alter initial beliefs about how predictable the event in question is, on the premise that less predictable events are more likely to be mentioned. We predict, consistent with the first and third scenario outlined above, that informationally redundant event descriptions should generate what we term atypicality inferences, where comprehenders resolve the apparent dip in informational utility by concluding that the usually predictable event described is in fact relatively atypical, as this would justify its mention. For example, in our scenario involving shopping and paying the cashier, a possible inference would be that ‘John’ does not pay predictably (e.g., has someone else pay for him, is a habitual shoplifter, or gets free groceries). We first look at these utterances in discourse contexts which implicitly support an atypicality inference, through semantically vacuous prosodic or discourse markers. 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, triggers weaker atypicality inferences, consistent with the third scenario, and models which explicitly take into account the influence of context on the likelihood of inference generation.

In the next section we review relevant literature, and in the following sections we describe the three experiments.

2 Related Work

Our work builds on two primary strains of research: interpretation and perception of informational redundancy on the one hand, and relatively new work on inferences about background world states (vs. speaker intentions) on the other. We also look at the effects of implicit prosody on pragmatic interpretation, which to date has largely been limited to semantic effects of contrastive prosody. Further, we look at the (systematic) interpretation of particularized, or ad-hoc pragmatic inferences, which deviate from communicative norms only in context, and/or on the basis of reasoning about world knowledge. These have not received a lot of attention from pragmaticists, experimentally or otherwise, partially due to their idiosyncratic nature.

2 When communicating any given pragmatic meaning, speakers frequently use multiple cues consistent with the intended meaning, in order to make the inference maximally easy to calculate for the listener (consider, for example, the difficulty of interpreting sarcasm without supportive prosody). We therefore consider the presence of supportive prosodic markers, or other supportive context, to be the ‘default’ case.
making them difficult to study systematically; and partially due to being seen as
less relevant to a theory of pragmatic vs. semantic meaning than scalar implicatures
(Levinson 2000).

2.1 Informational Redundancy

First, we want to discuss a problem of terminology. In most experimental literature,
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, overinformativeness has at
times been puzzlingly used to refer to both informational redundancy (Engelhardt
et al. 2006, Grice 1975), as well as to the relative informativeness of terms in an
implicational scale (e.g., the use of some when all is sufficient) (Levinson 2000).
The latter variety of overinformativeness, also typically associated with the Quantity
Maxim, is more a problem of vagueness where a more precise description is avail-
able. Informational redundancy, in contrast, is a problem of excessive wordiness or
precision, 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 con-
text), 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, and in the
literature on nominal overspecification.

Although informationally redundant utterances are typically regarded as infelici-
tous in the linguistics literature, they have been observed to be surprisingly common
in natural dialog. Baker et al. (2008) observed that such utterances are frequently
used in response to signs of listener non-comprehension, when responding to listener
questions, or when speaking to strangers. Walker (1993) also concludes that informa-
tionally redundant utterances are specifically used to address cognitive resource
limitations (e.g., memory for preceding discourse, limited inference-making capac-
ity), as well as to serve a narrative function. In the latter case, this may for example
involve drawing attention to a particularly salient or relevant fact. In other words,
many informationally redundant utterances are not really redundant, as any apparent
redundancy has communicative purpose. Literature on nominal overspecification
has similarly found that speakers are extremely likely to attach ‘redundant’ color
descriptions to nouns, even when doing so provides no new information regarding
which object is being referred to. However, in this case as well, there is evidence
that most overinformative nominal modification is not in fact overinformative, as
‘overdescribing’ an object can facilitate rapid and efficient object identification.
Here we will review some of the experimental work on informational redundancy, with a focus on interpretation of nominal overspecification.

Most experimental work on production and comprehension of informationally redundant utterances has focused on nominal modification in referent identification tasks, which typically instruct participants to look at or somehow engage with items such as: the [red] apple, the [tall] boot (Engelhardt et al. 2006, Nadig & Sedivy 2002, Sedivy 2003, Davies & Katsos 2010, 2013, Pogue et al. 2016). The aim of these studies has been to determine some combination of the following: 1) whether overinformative descriptions are perceived as infelicitous by listeners (i.e., whether overinformativeness violates communicative norms); 2) whether overinformativeness helps, hinders, or has no effect on object identification; 3) whether listeners attempt to accommodate overinformative descriptions by making inferences which increase the informational utility of the descriptions; and 4) whether listeners alter their beliefs about the rationality of the speaker (or the baseline informativeness of their speech) following use of overinformative descriptions.

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 with no other cups of any color) (Engelhardt et al. 2006, Nadig & Sedivy 2002 : 30% and 50% of nominal descriptions were overspecified in spontaneous speech, respectively). Further, listeners frequently do not find such utterances infelicitous: Engelhardt et al. (2006) showed that comprehenders judge overinformative descriptions as significantly more acceptable than underinformative descriptions, and that overinformative descriptions do not trigger additional (e.g., contrastive) inferences. Davies & Katsos (2010) find that overinformative expressions are more likely to be produced, and less likely to be judged infelicitous, than underinformative expressions, although they are still judged to be suboptimal\(^3\). Sedivy (2003) however showed that when listeners hear an object described with a clearly overinformative color adjective (e.g., “yellow banana”), they rapidly infer that a non-yellow banana must also be present.

What seems to emerge is that overinformative descriptions are easily tolerated when they describe perceptually useful or non-canonical properties, which may speed up object identification; and are more likely to be judged suboptimal, and/or trigger pragmatic inferences, when they don’t. Indeed, Rubio-Fernández (2016) showed that experimentally increasing the perceptual usefulness of color adjectives causes them to be produced more frequently, as well as that color adjectives are more likely to be used for atypical than typical colors. In a related line of research,

---

\(^3\) However, Davies & Katsos purposefully use adjectives less likely to be produced spontaneously - color adjectives, by far the most likely to be used redundantly, are avoided, and the adjectives that they use are largely either inherently contrastive (e.g., ‘tall,’ ‘big’); or describe a default, assumed state (e.g., ‘unbroken egg,’ ‘fresh apple’).
Pogue et al. (2016) found that after being exposed to a speaker repeatedly using overinformative (color or scalar) object descriptions, listeners are less likely to make generalization about the speaker’s rationality or informativity than when they use underinformative description. This suggests that listeners are relatively insensitive to deviations from “optimal” informativity that do not interfere with comprehension, or else perceive them as relatively commonplace and inconsequential.

Overall, this work has shown that some types of informational redundancy may be helpful to the listener, and that informational redundancy in general is tolerated by listeners. There is, however, also evidence that informationally redundant utterances that have no apparent (e.g., perceptual) utility are unlikely to be produced, are generally judged to be relatively infelicitous, and tend to generate inferences. 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 sense of not having communicative utility). Additionally, these studies are limited by the fact that they uniformly focus on a very particular, relatively concise, and lexically bounded variety of informational redundancy, raising the question of to what degree it’s possible to generalize from the results. What this points towards is a need to look at informational redundancy in the context of utterances and constructions that have no readily apparent utility to listeners - either in terms of comprehension, or in terms of facilitating the completion of a task. Further, we would argue that it’s 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 any existing norm against redundancy - for example, complex and lengthy multi-word utterances such as those in Example (2).

2.2 Common Ground Assumptions

To date there has been relatively little work on different strategies listeners might employ in interpreting an apparent violation (exploitation) of a conversational maxim. Most work has focused on the scenario where a listener detects an apparent maxim violation, assumes that a speaker is being cooperative, and comes up with an additional, non-literal meaning the speaker may have intended (which repairs the apparent violation). Another strategy is simply to assume that the speaker is being plainly uncooperative, or to assume that there is an intended meaning but that the listener is not privy to it, when no plausible intended meaning can be computed. A third strategy, which has received little attention, is that of modifying background assumptions when doing so would repair the violation, partially due to a focus on implicatures (intended meanings). To our knowledge, the only work to look at this in depth is Degen et al. (2015), which 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: listeners 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. Further, a prior belief that a scenario is strange significantly increases the strength of the some, but not not all implicature that is then drawn by the listener.

In our case, if, as in Example (1), a speaker states that John, having gone shopping, paid the cashier, a listener might ‘repair’ the redundancy by assuming that John does not typically pay the cashier. While this may occur parallel to an assumption that the speaker intended to use this utterance to communicate that John is a cashier-paying individual, the strategy of modifying background assumptions can well proceed without any assumptions about speaker intent. Perhaps the listener is a third party not privy to the background knowledge of the speaker and intended listener, or perhaps the speaker isn’t aware that the listener isn’t familiar with John’s paying habits. In fact, in the case of our example, it seems relatively unlikely that a speaker would choose to communicate information about John’s paying habits in this particular manner, making this an inference, but not an implicature. While most theoretical interest lies in implicatures, it’s important to be able to model prior and changing assumptions about the utterance common ground, given that they have been demonstrated to have a marked effect on which inferences are drawn by comprehenders, as well as their strength (Degen et al. (2015); see also the literature on presupposition, e.g.: Stalnaker (1973)). In our paper, we explore a method for testing the shifting of background assumptions, collect data that can be used in the future to test formal models of pragmatic reasoning, and explore the willingness of comprehenders to shift background assumptions in different contexts.

2.3 Effect of Implicit Prosody on Pragmatic Interpretation

One of the questions we ask, relevant to both accurately detecting and modeling an effect of information redundancy, is to what degree contextual cues (without semantic content) influence the detection and interpretation of informational redundancy. Although it is generally accepted that prosody may influence interpretation, there is very little empirical evidence outside the (possibly semantic) effect of contrastive prosody (Kurumada et al. 2012) that semantically empty prosodic changes to utterances have a substantial effect on implicature generation. One can however imagine that a redundant statement made loudly and confidently may cue a listener belief that the speaker is very intentionally communicating that particular bit of information to them, and that it should be taken seriously (signifying either that the
speaker is blatantly uncooperative by breaking a communicative norm for no reason, or that there is an additional reason that the information was transmitted). On the other hand, if a speaker vaguely mumbles an informationally redundant utterance under their breath, the listener might simply conclude that the speaker is reminding themselves of something, is unsure about what they really want to say, is mentally rehearsing a course of events, having some production difficulty, etc..

Along these lines, Bergen & Goodman (2014) hypothesize on the basis of formal pragmatic reasoning models that, rather than focal (contrastive) stress carrying conventional semantic meaning, the contrastive or exhaustive interpretation arises due to the listener perceiving the speaker as having made extra effort to communicate that particular bit of info to them. They argue that an utterance increased in volume or duration is more likely to be attended to and accurately perceived by the listener and that, conversely, speakers can intentionally exploit this to signal to listeners that such an utterance is important and specifically meant to not be confused with any alternative utterances. On the basis of this and similar work, we therefore experiment with having participants interpret both a bare informationally redundant utterance with (arguably natural) implicit exclamatory prosody (ending with an exclamation mark), as well as an informationally redundant utterance with no implicit prosody (ending simply with a full stop).

2.4 Context-dependent Implicatures

To date, most formal or experimental research on pragmatic inferences has focused on the use and interpretation of scalar implicatures (Horn 1984), such as the use of some to implicate not all, or warm to implicate not hot. Non-scalar ad-hoc inferences, which deviate from communicative norms only in context, and/or on the basis of reasoning about world knowledge, have not received much attention from pragmaticists, experimentally or otherwise. Traditionally, scalar implicatures have been regarded as a separate class of conventionalized inferences which rely minimally on context or general reasoning about speaker intentions (Levinson 2000), and which are based on specific lexical items (or classes of lexical items). In recent years this view has increasingly been challenged (Degen & Tanenhaus 2016, Grodner et al. 2010), with evidence indicating that the hypothesized differences between conventionalized (generalized) inferences, and particularized (ad-hoc) inferences, are in any case not categorical, although the nature of any differences remains difficult to determine, and the latter case has traditionally been understudied.

Research on conventionalized inferences has been critical to developing formal linguistic theory, due to the role they play in disambiguating pragmatic and semantic contributions to utterance meaning. However, we argue that context-dependent (ad-hoc) inferences, which occur far more frequently and ubiquitously, may be more
important to developing a more general theory of human communication, as originally intended by Grice (1975). There is an increasing, but still relatively small, body of experimental work teasing apart which properties of utterances trigger pragmatic inferences; what the sufficient or necessary criteria for triggering such inferences are; or which contextual cues speakers and listeners exploit in getting their intended meaning across, and interpreting intended meaning, respectively. 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). We therefore argue that further work on ad-hoc inferences on the one hand, and informational redundancy on the other, is important both for developing comprehensive and predictive theories of communication, and for developing formal models of pragmatic reasoning.

3 Experimental Procedure

The following experiments were run using the same interface, and on the same population of Amazon Mechanical Turk workers, in 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 3 ‘experiments’ was the manipulation of discourse markers on the utterance of interest 4. Running them concurrently and 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.

The number of eligible participants (n=2100) was predetermined through a simulation power analysis (adapted from Arnold et al. 2011): all predicted higher-order interactions, assuming effect sizes determined by a previous study (see Appendix A), were detectable at > .80. The R code and a plot for the power analysis can be found in the permanent online repository holding data and analyses for this article 5.

4 Experiment 1: Implicit Prosodic Emphasis

We first test whether informationally redundant event descriptions trigger atypicality inferences when there is minor contextual support for the implicature, in the form of an exclamation mark at the end of the utterance. Intuitively, exclamatory intonation is a natural way of framing a description of something ‘atypical’ (Rett 2011),

---

4 We did not make the manipulation of discourse markers a within-subjects factor, as that would significantly decrease the amount of within-experiment data we could collect from each participant, and, correspondingly, significantly decrease power and increase cost.

5 http://osf.io/kgv2r/?view_only=ff5859d3f33b485d95254395f95a52dc
without otherwise altering the semantic content of the utterance. When implicatures are context-dependent (and even if they are not; see Degen & Tanenhaus 2015), speakers generally provide multiple signals of their intended meaning, in order to make the implicature easier to calculate for the listener. One would expect this to particularly be the case when the intended meaning of the utterance substantially violates expectations or previously held beliefs, as opposed to simply providing new but marginally expected (or at least unsurprising) information.

We present naive participants with a limited number of brief ‘narratives,’ which set up the common ground context, relationships between discourse participants, and some typical or atypical properties of their normal behavior (where relevant). Some of the narratives include brief dialogue between two discourse participants at the end (which may include informationally redundant or informative event descriptions). After reading the narratives, participants rate how typical they believe certain behaviors in the story to be. We expect participants who read informationally redundant event descriptions to infer that the utterance in fact signals that the event is relatively unexpected, or atypical (as only relatively unexpected events warrant explicit mention). In contrast, those participants who read concisely informative event descriptions should draw no such inferences.

4.1 Methods

Participants

700 eligible participants (760 total; median age bracket 26-35; 50% female), were recruited on Amazon Mechanical Turk. The task was open only to workers located in the US, and with an approval rating of $\geq 95\%$. 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.

Those who did not provide accurate or plausible responses to the trial questions, all of which had a range of ‘valid’ and ‘invalid’ responses, were unable to proceed to the main task, and their data was not recorded (e.g., those who rated the likelihood of 50% heads on multiple fair coin flips as low, compared to other possible outcomes). Participants were likewise unable to proceed in the study, or submit their results, without having answered all questions.
Design

The primary question of interest is whether informationally redundant utterances (in this case, descriptions of highly predictable events) are perceived as deviating from conversational norms at face value (if interpreted literally), and whether they consequently trigger pragmatic inferences. These inferences should revise common-ground beliefs about the typicality of said events (and so ‘repair’ the violation, or dip in informational utility):

(3) “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 infer that John does not typically pay the cashier, as such a scenario would justify overt mention of John’s cashier-paying. The informational redundancy arises due to the high conceptual (or event) predictability of paying the cashier, and is resolved if one assumes that the event is not as predictable as initially assumed.

We also wanted to check whether the inference (that an event is less predictable than would otherwise be expected) could be cancelled by manipulating the common ground.

The event described becomes ‘non-predictable’ given a wonky common ground\(^6\) such as in (4), where the context suggests that typical assumptions (e.g., that one pays the cashier when they go to the grocery store) may not hold. At that point, the 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) 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 ‘concisely’ informative (non-redundant) event descriptions; and to confirm that similarly structured descriptions of conventionally non-predictable events as in ??, do not provoke similar inferences (which would suggest a problem with the stimulus design or response measure). In (4), the utterance is not informationally redundant, and is not expected to generate any specific inferences. We also wanted to confirm that the

\(^6\) 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.
wonky common ground inserted into the previous example does not affect the interpretation of conventionally non-predictable event mentions (which would suggest that there is an unexpected effect of context manipulation on stimulus interpretation, in general):

(5) **CONTEXT**: Ordinary or John habitually doesn’t pay.

“John just came back from the grocery store. He got some apples!”

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

**Materials**

24 stimuli were constructed as brief stories/narratives, based on distinct stereotyped scripts or activities. Each story had one of 2 context types (ordinary vs. wonky common ground, relative to the conventionally predictable or inferable script event). In all stories, declarative utterances, spoken by one of the discourse participants, described one of 2 types of script events (conventionally predictable vs. non-predictable), making a total of 4 conditions.

Conventionally predictable events (example (6)) can be inferred simply from the ‘speaker’ having invoked the script, while conventionally non-predictable events (example (7)) can not be inferred automatically, as they may only occasionally occur as part of the script activity. To clarify, we are using the term conventionally predictable to specify that the event almost invariably occurs as part of the event script (under normal conditions). Initial common ground was either ordinary ([1a], in (6)) with respect to the script, or wonky, in that it implied the conventionally or script-predictable event was in fact unusual for the event participant ([1b] below):

(6) **CONVENTIONALLY PREDICTABLE EVENT**

<table>
<thead>
<tr>
<th>[1a] John often goes to the grocery store around the corner from his apartment ordinary</th>
<th>[1b] John is typically broke, and doesn’t usually pay when he goes to the grocery store wonky</th>
</tr>
</thead>
</table>

[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.

---

7 The complete list of stimuli can be found in our online repository: [http://osf.io/49h5a/?view_only=ff5859d3f33b485d9525439f95a52dc](http://osf.io/49h5a/?view_only=ff5859d3f33b485d9525439f95a52dc)

The context/common ground manipulation in [1b] was used in order to render the conventionally predictable event unusual, or at least not predictable. Conventionally non-predictable events could not be automatically inferred from the script having been invoked:

(7) Non-predictable Event

[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.


Participants saw either only the common ground context [1] and discourse setup [2] (without numbering or special formatting), which enabled us to collect estimates of how typical activities are believed to be, based on the context alone (prior beliefs); or the entire text, which enabled us to collect estimates of how typical activities are believed to be, based on both the context and the event description [4] (updated beliefs).

Following each passage, participants were queried as to how typical they believed the conventionally predictable and non-predictable activities (as well as 2 other scenario-relevant distractor activities) were, for the person who is the subject of the discourse (the individual mentioned in the context [1] and event description [4]):

i. How often do you think John usually pays the cashier, when going shopping?
ii. How often do you think John usually gets apples, when going shopping?
iii. How often do you think John usually goes to the grocery store?
iv. 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 Figure 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.

12 of the stimuli included 3 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 12 only included two – 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 subac-
tivities or subroutines, and there needed to be predictable as well as non-predictable
subactivities (paying the cashier, getting apples). 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 ex-
pected to agree on what a script entailed, which activities were or weren’t obligatory
to the script sequence, etc..

All stimuli were normed on 3 qualities (in separate tasks): whether the activity
fell into the predictable or non-predictable activity bin; whether the common ground
manipulation was effective; and whether participants found it plausible that the
script could be engaged in without the predictable activity. For activity predictability
norming, participants were asked to rate the typicality of the activity (on a 0-
100 scale), with an arbitrary cutoff of 70 between activity types. Non-predictable
activities were on average rated 48.0 (25.1-68.1), and predictable activities were
rated 87.8 (78.1-95.2). For common ground norming, participants rated predictable
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 44.2; [19.8-72.9]); non-predictable 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 (predictable: 91% [67%-100%]; non-predictable: 94% [80%-100%]).
Measures

To measure comprehender beliefs regarding activity typicality, each story we presented was followed by 4 questions presented in random order, regarding activities mentioned in the story (including both conventionally predictable and non-predictable activities associated with the stimulus item). The questions were accompanied by sliding scales which ranged from Never to Always, where participants could select any point along the scale:

![Figure 1](image.png)

Figure 1   A sliding scale as used in the experiments.

Prior to seeing any experimental items, participants were given several practice questions, unrelated to the experimental stimuli, which also used continuous sliding scales ranging from Never to Always (or similar). 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 judged reasonably accurate, they were asked to re-read the instructions, and respond again, before they were able to proceed. This ensured that they were able to follow instructions, and were less likely to guess randomly throughout the experiment. There were no ‘accurate’ answers in the experiment itself. All points on the response scale were associated with a number ranging from 0 (Never) to 100 (Always).

Prior beliefs, or baseline beliefs regarding activity typicality, were estimated from responses to stimuli presented without the potentially redundant event description, or where the event description/utterance was not visible (see the next section for a more detailed explanation). The responses, aside from setting baseline measures (prior beliefs) of activity typicality, also provide a 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 highly inferable, within a given class, can be compared against one another.

Updated beliefs regarding activity typicality were estimated from responses to stimuli which included the redundant or informative utterance (event description), or where the event description/utterance was visible.

Belief change due to reading the event description was determined by modeling the magnitude and direction of difference between prior beliefs and updated beliefs.
Procedure

Participants were asked to read 6 experimental stimuli randomly selected out of the total of 24, as well as 4 filler items. 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 prior beliefs regarding activity typicality, and the 4 full stories gave us measures of updated beliefs conditioned on the event description.

<table>
<thead>
<tr>
<th>SUBJECT 1: prior belief</th>
<th>SUBJECT 2 updated belief</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;context&gt;</td>
<td>&lt;context&gt;</td>
</tr>
<tr>
<td>&lt;setting up of common ground&gt;</td>
<td>&lt;setting up of common ground&gt;</td>
</tr>
<tr>
<td>#. &lt;typicality question&gt;</td>
<td>#. &lt;typicality question&gt;</td>
</tr>
</tbody>
</table>

The experiment thus employed a between-subject design for belief measures, where prior and updated belief estimates for any given item were provided by different participants, to eliminate the possibility of participants conditioning their updated estimates not only on inferences made from the text, but also on their own prior estimates. The 4 filler stimuli had the same structure as above, but with the dialogue portion between quotes 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?”.

4.2 Results

For the purposes of determining whether participants made any inferences regarding activity typicality, we modeled belief change, i.e. the difference between prior and updated beliefs, or activity typicality estimates made with and without seeing the event description. Conventionally predictable and non-predictable activities were modeled separately, as the conventionally non-predictable activity was used primarily as a control, and manipulations of common ground context did not otherwise target it. All factors were effect/sum coded.

---

8 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.

9 However, the results of largely mirror the results of a within-subjects version of the study reported in Kravtchenko & Demberg (2015).
Conventionally Predictable Activities  ('Paid the Cashier')

Prior belief ratings (obtained from participants who did not see the event descriptions) showed that ordinary context events are perceived as highly typical (85.79 on a 0-100 scale). As predicted, updated belief ratings (obtained from participants who saw the here, redundant, event descriptions) show lower typicality for the ordinary context events (72.37) than prior belief ratings.

Wonky context events (i.e., the condition where the conventionally predictable event was made non-predictable by the common ground context) are perceived as relatively atypical a priori (48), and there was little change in participants’ ratings (45.71 for updated beliefs). The results are illustrated in Figure 2, using violin plots.

A linear mixed effects regression analysis 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 typicality 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 (prior / updated), as well as by-item random intercepts and slopes for both factors and their interaction (Barr et al. 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 et al. 2017).

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE($\beta$)</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>63.03</td>
<td>1.84</td>
<td>34.32</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Common Ground: Ordinary</td>
<td>32.38</td>
<td>3.33</td>
<td>9.72</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Belief: Updated</td>
<td>-7.83</td>
<td>1.71</td>
<td>-4.58</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Common Ground * Belief</td>
<td>-10.77</td>
<td>2.40</td>
<td>-4.50</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Table 4  Experiment 1: conventionally predictable activities 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.

These results show that, as predicted, when a conventionally predictable activity is seen in an ordinary common ground context (i.e. a context in which the activity can
Figure 2  Experiment 1: conventionally predictable activities analysis. This plot shows changes in activity typicality estimates depending on whether the utterance is present, as well as whether the context causes the utterance to be perceived as atypical. 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 activity frequency estimate. Circles represent mean values. Arrows show statistically significant differences between before/prior and after/updated ratings.

be automatically inferred), many readers infer that the conventionally predictable activity must in fact be non-predictable, i.e., atypical for the individual who is the subject of the story, and therefore worth mentioning explicitly.

Conventionally Non-predictable Activities  (‘Bought Some Apples’)

There was little change in participants’ ratings of conventionally non-predictable events from prior beliefs to updated beliefs (ordinary: 40.8 to 42.47; wonky: 38.49 to 39.56), see Figure 3.

A linear mixed effects regression analysis showed that estimates of activity typicality do not vary with the common ground context, nor are they conditioned on the utterance describing the activity (see table 5). This is also consistent with our predictions, and indicates both that the context alteration does not inherently cause a change in activity typicality estimates (regardless of how script-central the activity is), and that conventionally non-predictable activities, given our ordinary context, are not interpreted as (more) atypical when mentioned.
\[
\begin{array}{cccc}
  \beta & SE(\beta) & t & p \\
  \text{Intercept} & 40.29 & 1.86 & 21.69 & <.001 \\
  \text{Common Ground: Ordinary} & 2.88 & 2.07 & 1.39 & 0.2 \\
  \text{Belief: Updated} & 1.34 & 1.85 & 0.73 & 0.5 \\
  \text{Common Ground} \times \text{Belief} & 0.01 & 2.14 & 0.01 & 1 \\
\end{array}
\]

Table 5  Experiment 1: conventionally non-predictable activities analysis.

4.3 Discussion

The results of the first experiment indicate that comprehenders do in fact perceive informational redundancy, in the form of descriptions of overly predictable events, as a possible violation of conversational norms, and that they attempt to resolve the violation by reinterpreting the events described as atypical. On average, participants rate conceptually predictable activities as less typical if they see them mentioned overtly, in contrast to all other activities. These results provide evidence that informational redundancy, taken at face value, violates conversational norms, and that comprehenders react to it as they typically do to apparent maxim violations – by assuming an implied non-literal meaning that resolves the violation. This runs in some contradiction to Grice’s initial ambivalence about the existence of such a constraint, and equivocal evidence from studies of informationally redundant nominal modification.
These results rule out the second hypothesis outlined in Section 1.2, and raise two questions that we address in the following experiments, regarding the importance of (implicit) prosody, and that of contextual support for the inference. First, an exclamation point may serve multiple purposes: it may signal surprise as to the course of described events, a speaker’s intentionality in communicating a piece of information\textsuperscript{10}, the importance and relevance of the information conveyed to the general discourse and listener’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\textsuperscript{11}. Although it could be argued that the exclamation point (often a signal of surprise, Rett 2011) forces the ‘atypical event’ interpretation independent of utterance informativity, this is not a tenable explanation, as no signs of a similar effect are present in any of the other event descriptions.

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 remarkably little work on the question of which contextual cues specifically people employ in computing context-dependent inferences, as well as how these cues influence final pragmatic interpretation. To test this, in Experiment 2 we use a discourse marker (\textit{Oh yeah, and...}) which does not explicitly signal surprise, in order to frame the event description as intentionally conveyed, 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 listener, to lower the time and effort needed to get a message across. A charitable listener may be expected to expend considerable effort on rescuing the assumption of a cooperative or rational speaker (Davidson 1974). If only events under a certain threshold of predictability deserve mention (assuming a rational speaker), then listeners should conclude that the event mentioned is relatively atypical.

On the other hand, implicatures must be calculable (Levinson 2000) – that is, particularly for non-generalized (context-sensitive) implicatures, the context must offer sufficient support that the reader can guess the speaker’s intent with reasonable certainty. Likewise, while rational speakers may only mention events that are not easily inferable, forcing a listener to expend significant effort on recovering an utterance’s intended meaning is arguably non-rational behavior itself. In addition,

\textsuperscript{10} I.e., the speakers displays clear and conscious intent to draw to the listener’s attention 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.

\textsuperscript{11} For example: \textit{He paid the cashier. Then he noticed it was his classmate!}
the degree of “intentionality” on the part of the speaker (also signaled in our stimuli by the exclamation mark) may 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. To test whether informational redundancy itself is sufficient for triggering the inference, or whether some amount of discourse or prosodic contextual support is necessary for its generation, in Experiment 3 we present readers with the same task and stimuli, but strip the event description of prosodic or discourse cues supporting the inference.

5 Experiment 2: Implicit Discourse Support

The second experiment tests whether the effect, of informationally redundant event descriptions being interpreted by readers as signaling event atypicality, is generalizable. To do so, we 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 listener, 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 event having occurred, nor 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.

5.1 Methods

Participants

700 eligible participants (787 total; median age bracket 26-35; 51.3% 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.

Design

The design of this experiment was motivated by the same factors as Experiment 1—with the exception of how the event description was framed. Instead of marking the target utterance with an exclamation mark, we framed the same utterance as a piece of information the speaker had just recalled, apparently having previously intended to mention it to the listener:
“John just came back from the grocery store. **Oh yeah, and he paid the cashier.**

The *oh yeah...* discourse marker does not conventionally signal surprise, and therefore does not specifically signal the inference that we are testing for. It does, however, imply speaker intent behind conveying precisely this message, the importance and relevance of the message to the current discourse and listener, as well as that the message is able to stand alone, and is not intended to simply serve as causal or temporal scaffolding for a further message/event.

**Materials**

The same 24 stimuli were used as in Exp. 1. In this case, the critical utterance was prepended by *Oh yeah, and...:*:

(9) **ORDINARY CONTEXT**

1. John often goes to the grocery store around the corner from his apartment.
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** predictable.**”
4. **4b Oh yeah, and he got some apples** non-predictable.”

**Procedure**

The procedure was identical to that of Exp. 1.

**Measures**

The same response measures as in Exp. 1 were used to estimate *prior beliefs* and *updated beliefs*.

**5.2 Results**

As in Experiment 1, to determine whether participants made inferences regarding activity typicality, we modeled *belief change* - the difference between *prior* and
updated beliefs. Conventionally predictable and conventionally non-predictable activities were again modeled separately. All factors were effect/sum coded.

**Conventionally Predictable Activities**

As we predicted, prior belief ratings for ordinary context events showed that these events are judged to be highly typical (84.71). As in Experiment 1, updated beliefs about the typicality of ordinary context events were significantly lower (73.84), and wonky common ground estimates remained stable (47.45 prior to 47.47 updated).

A linear mixed effects regression analysis showed an interaction between context and belief measure ($\beta = -11.71$, $p < 0.001$), which is driven by lowered activity typicality ratings when the readers see the utterance in an 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 Figure 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 7.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE($\beta$)</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>63.58</td>
<td>1.85</td>
<td>34.33</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Common Ground: Ordinary</td>
<td>31.60</td>
<td>3.35</td>
<td>9.43</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Belief: Updated</td>
<td>-5.31</td>
<td>1.38</td>
<td>-3.83</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Common Ground * Belief</td>
<td>-11.71</td>
<td>2.03</td>
<td>-5.76</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

**Table 7** Experiment 2: conventionally predictable activities analysis.

These results support our prediction that, as long as the ‘atypicality’ inference has contextual support, in the form of implied speaker intentionality, message relevance, or message “encapsulation,” readers perceive the informationally redundant utterances as abnormal and make pragmatic inferences (of event atypicality), regardless of whether implicit prosody or other markers typically associated with surprisal are present.

**Conventionally Non-predictable Activities**

In contrast to Experiment 1, there was some increase in participants’ ratings of conventionally non-predictable events from prior beliefs (ordinary: 40.3 to 43.22; wonky: 37.74 to 43.05), see Figure 5.
A linear mixed effects regression analysis showed that estimates of activity typicality increase slightly when the utterance describing the conventionally non-predictable activity (see Table 8) 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-inferable, or concisely informative event may increase the perception of its typicality, 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.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE($\beta$)</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>40.99</td>
<td>1.85</td>
<td>22.14</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Common Ground: Ordinary</td>
<td>0.95</td>
<td>1.83</td>
<td>0.52</td>
<td>0.6</td>
</tr>
<tr>
<td>Belief: Updated</td>
<td>5.09</td>
<td>1.78</td>
<td>2.86</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Common Ground * Belief</td>
<td>−1.22</td>
<td>1.55</td>
<td>−0.79</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Table 8 Experiment 2: conventionally non-predictable activities analysis.
Figure 5 Experiment 2: conventionally non-predictable activities analysis.

5.3 Discussion

Together with Experiment 1, these results show that readers find conceptual informational redundancy abnormal, and will make pragmatic inferences to reconcile apparent informational redundancy with their expectations of utterance utility. This indicates that the effect is generalizable, and not dependent on conventional indicators of event atypicality, such as exclamatory intonation.

The results of Experiments 1 and 2 definitively rule out the second hypothesis described in Section 1.2, but do not permit us to distinguish between the 1st and 3rd hypotheses, as they leave open the question of whether the atypicality effect is dependent on context that supports the inference, or applies independently of discourse context. While some degree of contextual support is necessary for a (non-generalized) conversational implicature to be calculated (Levinson 2000), informational redundancy should be problematic in itself, and the implicit support offered by the implicit prosody and discourse marker in Experiments 1 and 2 is relatively minor. As discussed in the introduction, there is little evidence of semantically ‘vacuous’ prosodic changes to utterances having a substantial effect on implicature generation.

If the effect is dependent on discourse context that supports the inference, this would support the first hypothesis over the third, by suggesting one of the following. Comprehenders may be relatively unwilling to expend substantial effort on decoding an inference in the absence of supportive cues. Alternately, they may stop short in their efforts, on the assumption that it is more likely that speakers would violate expectations of informational utility, than that they would provide
insufficient contextual support for a comprehender to decode a non-literal meaning. Finally, they may 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 prosody or discourse markers that support the inference.

6 Experiment 3: Removing Contextual Support

To investigate whether implicit contextual support for the inference has an influence on the strength of the atypicality effect, we designed a third experiment which differs only in the absence of contextual support (in the form of implicit prosody or discourse marker) for the inference, or for the relevance/informativity of the event description. Our prediction is that while the effect may be attenuated, listeners should nevertheless make a measurable attempt to compensate for a violation in expected informational utility.

6.1 Methods

Participants

700 eligible participants (759 total; median age bracket 26-35; 51.6% 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.

Design

The design was motivated by the same factors as Experiments 1 and 2, but all implicit contextual support was removed from the event description:

(10) “John just came back from the grocery store. He paid the cashier.”

In this case, there is no specific support for the inference, or any signal of the relevance or informativity of the utterance. One could plausibly imagine the event description, in this case, to be ‘filler material,’ only semi-intentionally uttered while the speaker is planning what to say next, or as (planned, but then possibly abandoned) temporal or causal scaffolding for a more important event to be described, as in:

(11) “John just came back from the grocery store. He paid the cashier. He then realized he’d forgotten his driver’s license!”
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:

(12) ORDINARY CONTEXT

[1] John often goes to the grocery store around the corner from his apartment.
[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 predictable.”
[4b] He got some apples non-predictable.”

Procedure

The procedure was identical to that of previous experiments.

Measures

The same response measures as in the previous experiments were used to estimate prior beliefs and updated beliefs.

6.2 Results

As in previous experiments, we modeled the difference between prior and updated beliefs. Conventionally predictable and conventionally non-predictable activities were modeled separately. All factors were effect/sum coded.

Conventionally Predictable Activities

As in the previous experiments, prior belief ratings showed ordinary context events to be highly typical (85.59), and wonky context events to be less typical (49.5). Consistent with our predictions, updated beliefs are significantly lower in the ordinary context condition (80.3), but less so than in the previous two experiments. Exp. 3 is compared directly to Exp. 1 and 2 in Section 7.

A linear mixed effects regression analysis showed an interaction between context and belief measure (β = -5.4, p < 0.01), which is driven by lowered activity
typicality 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. A plot illustrating the interaction can be seen in Figure 6.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE($\beta$)</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>66.38</td>
<td>1.88</td>
<td>35.40</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Common Ground: Ordinary</td>
<td>33.21</td>
<td>3.40</td>
<td>9.77</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Belief: Updated</td>
<td>-2.20</td>
<td>0.93</td>
<td>-2.36</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Common Ground * Belief</td>
<td>-5.40</td>
<td>1.75</td>
<td>-3.10</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

**Table 10** Experiment 3: conventionally predictable activities analysis.

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 an ordinary common ground context, listeners do infer some degree of activity atypicality, even without implicit prosody or discourse markers serving as contextual support for the inference.
Conventionally Non-predictable Activities

In contrast to Experiment 1 and similar to Experiment 2, there was some increase in participants’ ratings of conventionally non-predictable events from prior to updated beliefs (ordinary: 41.08 to 46.46; wonky: 37.61 to 44.42), see Figure 7.

A linear mixed effects regression analysis showed that estimates of activity typicality do not vary with changes in the common ground context (or common ground wonkiness), but do increase slightly when the utterance describing the conventionally non-predictable activity (see Table 11) is visible ($\beta = 6.88$, $p < 0.001$). As in the case of Exp. 2, we suspect that explicitly mentioning a relatively unusual event causes participants to believe that event to be slightly more typical than they would otherwise assume.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE($\beta$)</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>42.12</td>
<td>2.12</td>
<td>19.84</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Common Ground: Ord.</td>
<td>2.29</td>
<td>2.41</td>
<td>0.95</td>
<td>0.4</td>
</tr>
<tr>
<td>Belief: Updated</td>
<td>6.88</td>
<td>1.77</td>
<td>3.88</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Common Ground * Belief</td>
<td>−1.39</td>
<td>1.72</td>
<td>−0.81</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Table 11  Experiment 3: conventionally non-predictable activities analysis.

Figure 7  Experiment 3: conventionally non-predictable activities analysis.
6.3 Discussion

In contrast to the results of the first two experiments, these results suggest that, when informationally redundant utterances are presented without a discourse context which implicitly supports the inference, comprehenders are relatively unlikely to draw atypicality inferences. This is consistent with the third hypothesis described in Section 1.2, and the premise that, while rational speakers may typically avoid making utterances that have no literal or implied informational utility, and while such utterances may prompt ‘compensatory’ pragmatic inferences on the part of listeners (which increase the informational utility of such utterances), such inferences must be sufficiently plausible in context and relatively easy for listeners to calculate (Levinson 2000).

7 Cross-Experiment Analysis and Gradience of the Typicality Effect

In this section, we directly compare the results of the three experiments. We predict that informationally redundant utterances trigger atypicality inferences independently of whether one uses an explicit marker of suprisal: in other words, that the effect is generalizable. However, we also predict that the effect is attenuated in absence of a discourse marker which signals relevance and speaker intentionality.

Conventionally Predictable Activities

To directly compare the three experiments, we run a $3 \times 2 \times 2$ linear mixed effects regression analysis of conventionally predictable activities. We modeled belief change (prior vs. updated 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 (‘.’ vs. the relevance markers).

The regression analysis showed a significant three-way interaction between discourse marker presence, common ground context, and belief measure: there was a significantly smaller atypicality effect in Exp. 3 than in Experiments 1 and 2 ($\beta = 5.78$, $p < 0.01$), and no significant difference between Experiments 1 and 2 ($\beta = -0.6$, $p = 0.8$).

We used the maximal converging model, with by-subject random intercepts and slopes for common ground context (ordinary / wonky) and belief measure (prior / updated), 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. A plot illustrating the comparison can be seen in Figure 8.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE($\beta$)</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>64.34</td>
<td>1.78</td>
<td>36.14</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>‘!’ vs. ‘Oh yeah...’</td>
<td>0.49</td>
<td>0.86</td>
<td>0.57</td>
<td>0.6</td>
</tr>
<tr>
<td>‘.’ vs. Relevance Markers</td>
<td>3.08</td>
<td>0.81</td>
<td>3.81</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Common Ground: Ordinary</td>
<td>32.40</td>
<td>3.22</td>
<td>10.05</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Belief: Updated</td>
<td>-5.07</td>
<td>1.04</td>
<td>-4.86</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>‘!’ vs. ‘Oh yeah’ * Common Ground</td>
<td>-0.64</td>
<td>1.43</td>
<td>-0.45</td>
<td>0.7</td>
</tr>
<tr>
<td>‘.’ vs. Relevance Markers * Common Ground</td>
<td>1.25</td>
<td>1.24</td>
<td>1.01</td>
<td>0.3</td>
</tr>
<tr>
<td>‘!’ vs. ‘Oh yeah’ * Belief</td>
<td>2.50</td>
<td>1.30</td>
<td>1.92</td>
<td>0.1</td>
</tr>
<tr>
<td>‘.’ vs. Relevance Markers * Belief</td>
<td>4.52</td>
<td>1.13</td>
<td>4.01</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Common Ground * Belief</td>
<td>-9.33</td>
<td>1.11</td>
<td>-8.41</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>‘!’ vs. ‘Oh yeah’ * CG * Belief</td>
<td>-0.60</td>
<td>2.35</td>
<td>-0.26</td>
<td>0.8</td>
</tr>
<tr>
<td>‘.’ vs. Relevance Markers * CG * Belief</td>
<td>5.78</td>
<td>2.03</td>
<td>2.84</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

Table 12 Experiments 1-3: conventionally predictable activities analysis.

As we predicted, the 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 context supporting the inference; in other words, participants are significantly less likely to make an atypicality inference in the absence of a discourse marker signaling relevance or intentionality.

The effect direction is consistent across items, with by-item common ground by belief measure interaction effect sizes ranging from -5.06 to -12.74. We further note here that this set of 3 experiments is a replication of a previously run set, as described in Appendix A. In addition, the ‘exclamation’ experiment in that section is a further replication of a within-subjects version (same stimuli), previously published as Kravtchenko & Demberg (2015), where participants updated their own ratings after seeing the utterance. We therefore argue that this is overall a robust and replicable effect.

12 A similar cross-experiment analysis of the conventionally non-predictable activities can be found in the supplementary materials: http://osf.io/2qdmj/?view_only=ff5859d3f33b485d95254395f95a52dc
This result clearly favors the third hypothesis described in Section 1.2 over the first. We conclude that in the absence of contextual support for the inference (a signal of clear speaker intentionality, the event description being an ‘encapsulated’ message, some indication of relevance to the listener/discourse), 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 rational conversational choices (Davidson 1974). However, it is also possible that comprehenders are less ‘charitable’ 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 strange or ‘irrational,’ 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 (11)). Determining which strategies comprehenders do in fact resort to, and in
which contexts, is left to future work. Finally, there is the possibility that, given the non-interactive experimental setting, comprehenders are processing the utterances at a relatively shallow level, and absent some (prosodic, discourse) indication that an utterance is somehow important, they do not expend effort on it (Sanford et al. 2006). To note, it has frequently been observed that comprehenders often do not make expected inferences in behavioral studies, for reasons that are not yet fully known (c.f., Noveck & Posada 2003). Determining whether this plays a role in our studies is left to future work, as is the question of whether similar or stronger effects can be observed in less artificial, and/or more interactive settings.

**Is the Effect of Typicality on Pragmatic Inferences Gradient?**

Figure 9 plots the average typicality estimate, before vs. after the target utterance, for each item in each condition, for all three experiments. The diagonal dashed line demonstrates what the null-hypothesis would predict: i.e., no effect of the utterance on belief change (prior ratings mapping straightforwardly onto updated 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 ground, conventionally-predictable 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 gradual ‘trend’ towards atypicality in the other three (non-redundant) conditions: items that are similar to ordinary predictable items, in terms of prior typicality estimates, are more likely to trigger atypicality inferences. In contrast, items with low prior typicality estimates show the opposite effect: i.e., if it’s mentioned that an individual engaged in a particularly unusual 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 gradual effect of prior beliefs in activity typicality on the likelihood of an atypicality inference, but this time the slope of 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 gradient difference between highly expected vs. optional 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, predictable
Figure 9 The plots show by-item belief change for all conditions of our three experiments. The dotted diagonal line represents the null hypothesis; i.e., what we would expect the data to look like if the critical utterance had no effect on beliefs. The solid black line is a regression line with 95% CIs across all conditions. The shading of the points represents the degree and direction of belief change: negative/black indicates an atypicality inference; positive/light gray indicates a perception of increased typicality.
activity event mentions. Furthermore, we can see that the effect of predictability on atypicality inferences is clearly gradient rather than binary: more predictable activities, in all conditions, generally elicit larger atypicality inferences.

8 General Discussion

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 suggests that listeners expect for speaker utterances to have a certain level of informational utility, and that they adjust their beliefs about the world and/or utterance meaning when such expectations are violated. In fact, this occurs even though informational redundancy, or overinformativity, in itself has no clear negative impact on comprehension. This is consistent with theoretic accounts of what constitutes ‘cooperative’ speaker behavior (Grice 1975), as well as of listeners’ attempts to resolve behavior that at face value does not appear rational or cooperative. However, as the third experiment shows, the effect is significantly modulated by how the utterance is framed in the discourse, supporting the third hypothesis outlined in Section 1.2. Overall, we provide robust evidence that informational redundancy is perceived as anomalous, and that comprehenders alter their situation models to accommodate it, particularly when context provides enough support for them to do so.

As discussed in Section 2, while redundancy may not directly negatively impair comprehension (unlike underinformativeness), comprehenders may nevertheless prefer or expect that speakers be relatively concise, as it allows them to receive more information in a shorter span of time. Excessive redundancy may make it more difficult to follow the point of a conversation, or to reliably distinguish important from unimportant information. Provided that speakers are aware of comprehender preferences, and are (possibly) independently motivated to conserve articulatory energy, it would be expected that listeners should perceive highly and unnecessarily redundant utterances (e.g., ‘yellow banana,’ ‘he went shopping and paid the cashier!’) as such. Correspondingly, they should infer that the speaker is either not behaving rationally13, or communicating a message or background world state that is unusual from the comprehender’s perspective. ‘Moderately’ redundant utterances, such as when a speaker points out ‘the long fork’ in the absence of another fork to compare, are not particularly useful in most tasks, but at least provide information that can’t otherwise be inferred from the rest of the utterance. We believe

13 See the section on Gricean vs. Bayesian rationality below: Section 8.1.1.
that the clearest results on how redundancy is perceived should come from utterances like the ones investigated here, or what we term highly redundant utterances.

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. The redundant utterance violates conversational norms, or listener expectations, under the default assumed world state, but not the alternate ‘wonky’ world state (e.g., one in which John is a habitual non-payer). Hearing such an utterance therefore biases comprehenders towards assuming that the alternate, or wonky, world state holds. Unlike shifting assumptions about intended utterance meaning, this is a strategy of accommodating potential violations of conversational norms 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, and neglecting it as a possibility likely results in misinterpretation of online effects and under-detection of pragmatic inferences in experimental work.

Finally, we show that semantically ‘vacuous’ utterance features (those that make no conventional contribution to meaning), in the form of implicit prosody or discourse markers, significantly influence the extent to which comprehenders are willing to draw an inference predicted by pragmatic theories of rational speaker behavior. Aside from the case of contrastive prosody (Bergen & Goodman 2014, Kurumada et al. 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 carefully weighing and evaluating multiple cues of how likely it is that a speaker intended to communicate a particular meaning, or that a deviation from expected utterance form and/or meaning signifies a common ground or background state that is substantially different from what was initially assumed.

### 8.1 Processing Difficulty and Surprisal

This subsection is primarily aimed at those who are interested in the processing cost of computing implicatures, or computational models of language processing. A question that we raise for future research is whether encountering informationally redundant utterances results in measurable processing difficulty on the part of comprehenders. We further argue that this has significant implications for current models of language processing.

As Walker (1993) points out, informationally redundant utterances are common in natural dialog - they therefore cannot be regarded as edge cases, and must be
integrated into our models of language. We believe, however, that they pose several unique challenges for formal models of language processing. There are several potential sources of processing difficulty associated with such utterances, resulting on the one hand from processing the surface form (the particular string of words that comprises the utterance), and on the other hand from computing the pragmatic inference itself\(^{14}\). First, there could be processing difficulty associated with the (un)predictability of the surface form of the utterance: John paid the cashier! is an utterance we would not expect to hear in a ordinary context, as paying the cashier is normal, and reading unpredictable utterances such as this should cause some difficulty (Smith & Levy 2013). Second, we work on the assumption that context-dependent implicatures incur processing cost (Levinson 2000, Sedivy 2007), although there is evidence that processing may be relatively rapid, provided the context adequately supports the inference (Degen & Tanenhaus 2015, Grodner et al. 2010).

8.1.1 Speaker Rationality

First, we want to briefly talk about the link between Gricean notions of rationality, and an information-theoretic or Bayesian approach to rational speaker behavior. Rationality in the Gricean sense concerns whether speakers are constructing their utterances in a manner that is consistent with their goals, which is accurately communicating their message to a listener (Grice 1975). To this end, underinformativeness (saying less than needed), for example, is clearly inconsistent with this goal. Saying more than needed, however, does not clearly impair one’s ability to accurately communicate a message - hence, the general uncertainty over whether overinformativeness violates Gricean norms. In the information-theoretic tradition, the speaker’s goal is to expend no more energy than needed to accurately transmit a message (Jaeger 2010). Expending more effort than required to accurately transmit a message is inefficient from the speaker’s perspective, and therefore not particularly rational, even while it is worse from a communication standpoint to not expend enough effort. The two traditions therefore make roughly similar predictions – weakly in the Gricean case, and more strongly in the information-theoretic: that redundancy should be avoided.

What the Gricean tradition adds to this mix is an idea of how listeners might interpret deviations from the communicative norm; traditional information-theoretic accounts make no predictions about how perceived utterance meaning might be altered when there’s a mismatch between the expected and actual utterance. However,

\(^{14}\) Although there is debate currently over just how rapidly or efficiently comprehenders are able to make pragmatic inferences, much of the evidence converges on there frequently being some cost, even for relatively conventionalized inferences (Degen & Tanenhaus 2016).
some recent accounts (Jaeger & Buz 2017) note that if the speaker’s aim is to accurately communicate a message, then they must take into account the signal, or *surface form*, that listeners expect to hear for that particular message. If they produce something that deviates from the expected signal, listeners may be led to assume a different intended message, which is more compatible with the signal that was in fact produced. With this development, perhaps, it might be possible to fully reconcile the various traditions, but we leave this to future work (see, however, Frank & Goodman 2012 for a formal model of how speakers’ and listeners’ reasoning about each others’ intentions might account for and predict utterance choice and interpretation, while incorporating cost-based considerations).

While the above covers rationality from the point of view of the speaker, on the comprehender side, information-theoretic models of language processing propose that listeners have strong expectations for how things will be said, and encounter processing difficulty (reflected by a variety of online measures) when these expectations are violated. In this tradition, what comprehenders specifically have expectations about is the *form* of utterances. After hearing something like *John went to the store*, they do not expect for *He paid the cashier!* to follow, as it is redundant: they are therefore surprised by the form the discourse has taken. The Gricean tradition similarly suggests that listeners have a base expectation that speakers will behave rationally, and interpret utterances literally or non-literally in a manner that will, generally, help to match this expectation (Grice 1975, Levinson 2000). We explicitly propose also that comprehenders have expectations as to the *state of the world* communicated by the speaker. In the above example, the state of the world is precisely the one that is expected (i.e., one in which *John* has paid the cashier). The two forms of predictability - *form-based* and *meaning-based* - are often treated as essentially identical, but as we discuss in the following section, need to be disentangled to make accurate predictions about language processing.

### 8.1.2 Surprisal

An area where our work might have particular implications is in formal modeling of language processing. The mathematical concept of *surprisal* (Hale 2001, Levy 2008), traditionally, represents how (un)predictable a word or a string of words is in context. Specifically, it is the negative log of the probability of encountering a specific word or utterance. As hinted in the name, words or utterances that one might expect to see in a given context have low surprisal values, and those that one would not expect to see in a given context have high surprisal values. Smith & Levy (2013) show that difficulty in processing a word (or string of words), as reflected in online measures like reading times, is proportional to the word’s unpredictability in context, or surprisal. In other words, comprehenders read or process words or utterances
that are predictable (low surprisal) quickly, and those that are unpredictable (high surprisal) slowly. An utterance you don’t expect to see in ordinary contexts (John paid the cashier!) should incur some processing difficulty for comprehenders. However, a problem with this account is that it treats all forms of predictability similarly. Consider, for example, two utterances that one might be (hypothetically) equally likely to hear: John paid the cashier! and John punched the cashier! Processing theories which take into account only the predictability of an utterance would predict similar processing difficulty or processing times for both. However, this is not only conceptually problematic, but would likely make the wrong predictions. Considering only surface-level or form-based predictability (the predictability of the string of words, in context) doesn’t take into account the fact that the utterances are unpredictable for entirely different reasons: dispreference for redundancy, vs. event unpredictability. Further, the first (cashier-paying) utterance may contribute additional processing cost due to encountering pragmatic abnormality, or due to the need to make a pragmatic inference to resolve the apparent redundancy. In this case, despite identical surface-form predictability, we would expect that conceptually redundant utterances would be associated with greater processing difficulty. Of course, it may also be the case that conceptually redundant utterances are relatively easy to process, due to the relative ease of semantic integration (there are no atypical or unexpected facts to integrate into one’s world model15). Either case, however, poses problems for the link between surprisal and processing difficulty, as utterances matched on predictability (and, consequently, surprisal) would still not end up with identical processing difficulty or reading times.

Several other interesting implications remain for surprisal theory, or the claimed link between surprisal, and reading times, or processing cost. First, it is commonly assumed that processing difficulty, in the context of this theory, is caused by encountering a particularly unexpected form. However, in our redundant cashier-paying example, the form of the utterance is unexpected precisely because the predictability of the utterance meaning is so high. In other words, in order for comprehenders to have expectations about the form of the utterance, they must also have expectations about the meaning, as it is precisely the meaning that renders the form surprising to comprehenders. We therefore consider it a significant shortcoming of these theories that they frequently either do not consider the predictability of meaning (what can also be termed conceptual predictability), or treat the two probabilities, form and meaning, as essentially identical, whereas we have shown that they not only can influence each other, but in fact can diverge systematically at their extreme values. In

15 Of course, our experiments make clear that many comprehenders do end up integrating an unusual common ground belief (John is a habitual non-payer) when trying to resolve the apparent pragmatic violation. For those comprehenders, we would predict that the processing cost would, in fact, be greater than the cost of simply integrating an unusual event into one’s world model.
the following subsection, intended for readers interested in language processing and formal language models, we talk about this relationship in more detail, as well as the implications it has for what types of language models could in principle address the issues we’ve outlined.

### 8.1.3 Formal Models of Language Processing

The predictability of informationally redundant utterances, as we’ve mentioned, should be fairly low at the *surface* level, and reading times have been argued to reflect the predictability of *surface-level* linguistic events, rather than the conceptual predictability of the scenarios they describe, i.e., their *meaning* (Smith & Levy 2013). There is evidence, however, that comprehenders predict at multiple levels: for example, the *event* (in our case, *meaning*) level, as well as at the level of *surface form* (Kuperberg & Jaeger 2016), although it remains unclear how these levels interface (e.g., if comprehenders expect something predictable at the *event* level to go unmentioned at the *surface* level). In the case of surprisal theory (Hale 2001, Levy 2008), this may have interesting implications, given that the surprisal values that have been linked to processing times have largely been obtained using formal (computational) language models. If informationally redundant utterances result in longer reading times, it’s unclear how formal models could accurately generate the high surprisal values one would expect for those utterances\(^\text{16}\).

For example, in the case of our *conventionally predictable* event utterance, in the *ordinary* vs. *wonky* common ground, the event description (*John paid the cashier*) consists of exactly the same string of words, with the preceding context identical stretching over multiple preceding sentences. The utterance is informationally redundant in the *ordinary* context, and what could be termed *concisely informative* in the *wonky* context. Simple or even complex n-gram models, which can’t represent long-distance dependencies, would not show any difference in predictability, and therefore would predict no differences in processing difficulty. Relatively sophisticated models which incorporate syntax or semantics, similarly, would not predict a difference, as there are no meaningful differences in syntactic structure, and semantic models would not have access to the relevant event-based information which distinguishes the utterances.

Models of event sequences, which estimate *event* (vs. *string*) probability, may be able to estimate differences in predictability, and, consequently, processing difficulty, between utterances describing script-congruent and script-incongruent events (e.g.,

\[\text{16 For that matter, if they result in shorter reading times, as speculated in the previous section, there would similarly be a problem given that the processing difficulty should not rely simply on *surface-level* probability, but also on *event* or *meaning* probability, which, as we explain, current models cannot adequately integrate.}\]
events likely and unlikely to be a part of grocery shopping). However, the general prediction such models would make is that the more congruent an event is with an invoked script (i.e., the more predictable the event is given the script), the more predictable (and easy to process) the utterances which describe that event should be. There is no principled way, within this framework, to divide events up into different ‘grades’ of predictability, such that utterances describing moderately predictable events are easier to process than those describing not-so-predictable events, yet those describing very predictable events incur difficulty. In light of this, we suggest that to predict any processing difference between informationally redundant and concisely informative utterances, formal models of language comprehension would need to incorporate some form of pragmatic reasoning.

Although attempts to build formal or computational language models may appear to have limited relevance to how humans process language - which is typically thought of as a seamless integration of information from the surrounding context - it should be recognized that humans do not make predictions about upcoming material based simply on the preceding string of words, as formally assumed by simple models of language processing and prediction. The vast majority of word/utterance sequences have never been previously encountered by a comprehender, and predictions concerning upcoming material cannot be based on them alone. Regardless of the modeling approach one takes, it must be concluded that humans also make predictions by keeping track of certain cues - semantic, syntactic, lexical, and pragmatic (e.g., whether a speaker is generally adhering to conversational norms). Thus, determining the specific cues that are necessary to accurately model language processing is also relevant to understanding how humans accomplish the same task, and what information they must keep track of in order to do so. There are two ways of elucidating which linguistic and contextual cues influence language comprehension: one may manipulate relevant cues in tightly controlled stimuli, and observe their influence on interpretation, or online measures such as reading times; or one may build formal language models which make specific, testable predictions regarding the influences of these cues on processing and comprehension. We believe that a combination of the two is likely to be the most fruitful approach.

To summarize, we think that it would be informative to investigate the processing of informationally redundant utterances, using online measures such as eye-tracking of self-paced reading. On the one hand, there are many claims, but still relatively little data on the online processing of pragmatic inferences, and little is known about the cost (or efficiency) of pragmatic reasoning (Degen & Tanenhaus 2016). The data that does exist is for the most part limited to scalar implicatures, which are often argued to be computed relatively automatically (but see Huang & Snedeker 2009). On the other hand, determining whether informational redundancy contributes to the processing cost of utterances, above and beyond the surface predictability of
those utterances, is critical to determining whether formal language models need to integrate pragmatic reasoning to correctly predict processing cost. The main challenge to using online measures is that to compare (for example) the reading times of utterances, they must be matched on all factors which may affect reading times, but are irrelevant to the experimental manipulation (in a case such as ours: length, word frequencies, etc.). This makes it very difficult to compare reading times for utterances that are not identical in their surface form. One possibility is to compare reading times for otherwise identical phrases that are informationally redundant in one context, but not the other, as with our cashier-paying examples in the ordinary and wonky common grounds. We leave this to future work.

8.2 Perspectives for Future Work and Conclusion

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 & 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’s 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.

Overall, our results show that, at least at face value, informational redundancy is perceived as anomalous. However, comprehenders are able to accommodate the provision of ‘unnecessary’ information by altering their prior beliefs about individuals’ behavior, or, more broadly, the common ground between speaker and listener. The results also complement work in the dialogue literature (Walker 1993), which illustrates that informationally redundant utterances are frequently used to convey ‘informationally useful’ non-literal content. They raise presently important questions regarding which cues are systematically tracked by comprehenders, as well as how those cues are integrated during pragmatic interpretation. Finally, they address the interpretation of complex, non-lexically-bounded utterances which to
date have largely been treated as either too complex or too idiosyncratic to study systematically.

References


A Replicated Work

In this appendix, we present a previous iteration of this series of experiments, using the same design as that reported in the main body of the paper, but run on separate populations (as opposed to concurrently), and with a slightly different set of stimuli. We include these results here as evidence that the effects we report are robust, replicating closely despite being run in a different population, substantial revision of the stimuli to improve naturalness, addition of filler stimuli, and a larger amount of data being collected to improve power for all relevant comparisons.

A.1 Methods

Participants

1200 eligible participants (1242 total), 400 per experiment, were recruited on Amazon Mechanical Turk, with the task only open to workers located in the US, and with an approval rating of $\geq 95\%$. Participants who did not report their native language, or reported their native language as other than English, were excluded (42; 3.38%), with additional participants recruited to replace them.

Materials

The design was identical to that reported in the paper, aside from the inclusion of fillers, as each participant saw only 6 stimuli and no condition more than once, with all stimuli differing across multiple non-critical dimensions. We therefore reasoned that there was little likelihood of learning the purpose of the experiment in the course of the task, and there was risk of increased task length/tedium decreasing the likelihood of participants reading passages closely enough to pick up on relatively subtle effects.

The stimuli in the replicated experiments were constructed to minimize variation in syntactic and information structure, as well as length, between stimuli. However, this came at the cost of naturalness. Here we present a stimulus example analogous to (6) and (7):
Procedure

The procedure was identical to that of the other experiments.

Measures

The same response measures as in the other experiments were used to estimate prior beliefs and updated beliefs.

A.2 Results

As in the experiments reported in the main body of the paper, we modeled the difference between prior and updated beliefs. Conventionally predictable and conventionally non-predictable activities were modeled separately. All binary factors were effect/sum coded, and the experiment factor was Helmert coded.

Conventionally Predictable Activities

The regression analysis showed a significant three-way interaction between discourse marker presence, common ground context, and belief measure: there was a significantly smaller atypicality effect in Exp. 3 than in Experiments 1 and 2.
(β = 6.01, p < 0.05), and no significant difference between Experiments 1 and 2 (β = 0.08, p = 0.98).

We used the maximal converging model, with by-subject random intercepts and slopes for common ground context (ordinary / wonky) and belief measure (prior / updated), as well as by-item random intercepts and slopes for all factors. By-subject random slopes for the interaction were not included in the model, because we did not have any repeated measures for subjects for the interaction. By-item random slopes for the interactions were not included in the model due to nonconvergence. A plot illustrating the higher-order experiment by common ground by belief measure interaction can be seen in Figure 10.

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE(β)</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>61.23</td>
<td>2.11</td>
<td>29.06</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>‘!’ vs. ‘Oh yeah...’</td>
<td>1.27</td>
<td>1.24</td>
<td>1.02</td>
<td>0.3</td>
</tr>
<tr>
<td>‘.’ vs. Relevance Markers</td>
<td>4.32</td>
<td>0.99</td>
<td>4.36</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Common Ground: Ordinary</td>
<td>38.04</td>
<td>3.72</td>
<td>10.24</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Belief: Updated</td>
<td>−0.43</td>
<td>1.43</td>
<td>−0.30</td>
<td>0.8</td>
</tr>
<tr>
<td>‘!’ vs. ‘Oh yeah’ * Common Ground</td>
<td>−0.86</td>
<td>1.82</td>
<td>−0.47</td>
<td>0.6</td>
</tr>
<tr>
<td>‘.’ vs. Relevance Markers * Common Ground</td>
<td>2.54</td>
<td>1.57</td>
<td>1.62</td>
<td>0.1</td>
</tr>
<tr>
<td>‘!’ vs. ‘Oh yeah’ * Belief</td>
<td>0.72</td>
<td>1.76</td>
<td>0.41</td>
<td>0.7</td>
</tr>
<tr>
<td>‘.’ vs. Relevance Markers * Belief</td>
<td>6.80</td>
<td>1.52</td>
<td>4.48</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Common Ground * Belief</td>
<td>−12.64</td>
<td>1.27</td>
<td>−9.96</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>‘!’ vs. ‘Oh yeah’ * CG * Belief</td>
<td>0.08</td>
<td>3.11</td>
<td>0.03</td>
<td>1</td>
</tr>
<tr>
<td>‘.’ vs. Relevance Markers * CG * Belief</td>
<td>6.01</td>
<td>2.69</td>
<td>2.23</td>
<td>&lt;.05</td>
</tr>
</tbody>
</table>

Table 14     Replicated Experiment 1-3: conventionally predictable activities analysis.

A similar analysis of the conventionally non-predictable activities can be found in the supplementary materials17.

A.3 Discussion

Overall, the results of these experiments were broadly replicated by those reported in the main body of the paper. The only salient difference is that in the original iteration of Exp. 3, there was no measurable effect of informational redundancy on perceptions of activity typicality, while in the ‘new’ Exp. 3, there was a significant,

17 Supplementary materials: http://osf.io/2qdmj/?view_only=ff5859d3f33b485d95254395f95a52dc
Figure 10  Replicated Experiments 1-3: conventionally predictable activities analysis.

but diminished effect, as we had originally predicted. The absence of a significant effect in the first iteration surprised us, and we attribute it to either chance (possibly due to fewer subjects run) or to increased prominence of the utterance in the revised stimuli. To compare:

(14)  REvised: “John just came back from the grocery store. He paid the cashier.”

(15)  ORIGINAL: “John went shopping. He paid the cashier. I just saw him in the living room.”

The utterance in question appears more discourse-prominent in the revised version of the stimuli, as it is utterance-final (i.e., we removed the last sentence), and in general competes with fewer adjacent utterance. We leave it to future work to definitively answer whether the minor change in utterance prominence does indeed eliminate the effect entirely.