Cognition 209 (2021) 104491

Contents lists available at ScienceDirect

Cognition

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

What's new? A comprehension bias in favor of informativity

Hannah Rohde [a,*], Richard Futrell [b], Christopher G. Lucas [c]

[a] Linguistics & English Language, University of Edinburgh, UK
[b] Language Science, University of California Irvine, USA
[c] Informatics, University of Edinburgh, UK

ARTICLE INFO

Keywords:
Language comprehension
Predictability
Pragmatics
Reading time

ABSTRACT

Language is used as a channel by which speakers convey, among other things, newsworthy and informative messages, i.e., content that is otherwise unpredictable to the comprehender. We therefore might expect comprehenders to show a preference for such messages. However, comprehension studies tend to emphasize the opposite: i.e., processing ease for situation-predictable content (e.g., *chopping carrots with a knife*). Comprehenders are known to deploy knowledge about situation plausibility during processing in fine-grained context-sensitive ways. Using self-paced reading, we test whether comprehenders can also deploy this knowledge in favor of newsworthy content to yield informativity-driven effects alongside, or instead of, plausibility-driven effects. We manipulate semantic context (unusual protagonists), syntactic construction (wh- clefts), and the communicative environment (text messages). Reading times (primarily sentence-finally) show facilitation for sentences containing newsworthy content (e.g., *chopping carrots with a shovel*), where the content is both unpredictable at the situation level because of its atypicality and also unpredictable at the word level because of the large number of atypical elements a speaker could potentially mention. Our studies are the first to show that informativity-driven effects are observable at all, and the results highlight the need for models that distinguish between comprehenders' estimate of content plausibility and their estimate of a speaker's decision to talk about that content.

1. Introduction

Comprehension can be said to involve, or at least approximate, a process of reverse engineering. Comprehenders attempt to reconstruct the underlying production process that might have given rise to the surface forms they encounter. Comprehenders are understood to have expectations about what messages a speaker could be trying to convey. They use these expectations to guess what content is coming next, such that content that is more expected is easier to process. The challenge then for modelling language comprehension is to characterize these expectations: Do comprehenders track word co-occurrence patterns? Do syntactic constraints matter? Is real-world knowledge deployed in real time? The answer to all of these has been shown to be yes; comprehenders bring to bear a remarkable number and sophisticated combination of cues during sentence processing. They show domain-specific sensitivity to statistical frequencies in their language input regarding the words and constructions speakers use (e.g., Gries & Divjak, 2012), alongside domain-general awareness of plausible events and situations that a speaker might be describing (e.g., Kutas & Hillyard, 1980).

It is the latter competence that is the focus here. However, in contrast to prior work on real-world knowledge that emphasizes comprehenders' ease in processing *situation-typical* material, here we consider a bias towards newsworthy, and hence *situation-atypical*, information. The goal is to test whether comprehenders can ever be shown to favor newsworthy messages (Grice, 1975). We use the term 'situation typicality' to refer to the probability of situations in the world and 'utterance expectedness' to refer to the felicity of an utterance as an appropriate contribution to a discourse. So comprehehenders may expect speakers to produce utterances about atypical (real-world implausible) situations (because such content would be discourse appropriate), even though it means they expect speakers to talk about content that is predictable neither at the situation level nor at the word level. A piece of newsworthy content may be interesting precisely because it is drawn from the infinite set of atypical situations, rendering any individual piece of informative news

* Corresponding author.
E-mail address: hannah.rohde@ed.ac.uk (H. Rohde).

https://doi.org/10.1016/j.cognition.2020.104491
Received 6 July 2019; Received in revised form 6 October 2020; Accepted 20 October 2020

very low probability.[1] As an illustration, imagine a speaker who looks into a room and announces one of the following:

(1)
a. The room is full of people.
b. The room is full of zombies.
c. The room is full of air.

The intuition is that (1-a) might constitute a fairly reasonable utterance — not all rooms contain people but some do, and a room full of them might be worth reporting. In contrast, (1-b) might be unexpected since zombies are rare in the real world. However, (1-c) might also be unexpected, not because the situation being described is rare, but because the situation is too ubiquitous to be worth mentioning. Fig. 1a draws a hypothetical probability distribution over situations — higher probabilities for rooms containing people or chairs or air; lower probabilities for rooms containing diamonds or zombies. Fig. 1b shows a distribution intended to capture the likelihood of a speaker choosing to report situations of different probabilities. A room containing air is highly probable so a speaker is unlikely to mention it. By the same token, a room containing zombies is improbable, but if a speaker were to encounter such a room, choosing to report it is highly likely, even if, as a listener, it is nearly impossible to anticipate which specific low-probability situation an informative speaker will mention.

These intuitions can be mapped onto formal measures of expected utility and costs of particular utterances (e.g., Benz, Jäger, & Rooij, 2005; Lewis, 1969). In Bayesian terms, situation typicality (Fig. 1a) represents the priors over situations; utterance production (Fig. 1b) corresponds to the likelihood of producing a surface form about different situations. Existing language processing models tend to emphasize one or the other. Models of comprehension focus on the prior, as measured in comprehenders' awareness of situation plausibility whereby a description of a typical situation is easier to process than a description of an atypical one (see review in McRae & Matsuki, 2009). On the other hand, models of production aim to capture the likelihood, specifically the observation that speakers omit words that are uninformative (Brown & Dell, 1987; Dale & Reiter, 1995), in keeping with information-theoretic approaches that link predictability to reduction (Aylett & Turk, 2004; Levy & Jaeger, 2007). In this sense, studies on production take the underlying content as given and focus on how that content is most likely to be realized, so utterance probability is modelled as the probability of the most likely surface form that a speaker would select in order to convey their meaning, as per (2).

(2) $p(\textit{utterance}) \propto \arg\max_i p(\text{form}_i|\textit{meaning})$

The probability of different meanings in turn is the purview of comprehension research, with abundant evidence of comprehenders' bias in favor of semantically plausible meanings (e.g., Gibson, Bergen, & Piantadosi, 2013; Hagoort, Hald, Bastiaansen, & Petersson, 2004; Kutas & Hillyard, 1980; McRae, Spivey-Knowlton, & Tanenhaus, 1998; Walker, 1975). Such studies make a tacit assumption that (listeners believe that) utterances describing more plausible meanings are the ones speakers are more likely to produce. If we characterize speakers as selecting candidate utterances directly from a probability distribution skewed towards the plausible, this characterizes language production as a transparent mapping from real-world situations to surface utterances, as in (3).

(3) $p(\textit{utterance}) \propto p(\textit{meaning})$

However, that mapping need not be transparent. Rather, language may reflect two components: the possibility that a speaker observes (or contemplates or remembers or wishes for, etc.) a particular situation and the likelihood of whether/how the speaker articulates that observation (or contemplation or memory or wish). Plausible meanings may be too mundane, but really newsworthy ones may be too rare. If we posit a generative architecture in which sampling from these two distributions is what yields speakers' articulation of particular utterances, production would look more like (4), where the prior over meanings is combined with the likelihood of producing a particular surface form.

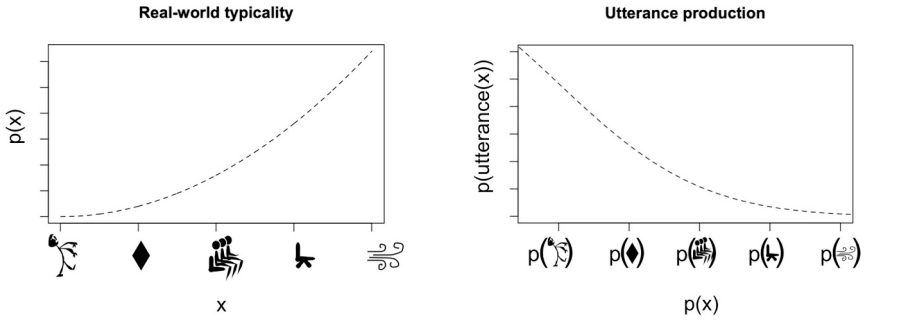
(4)
$$p(\textit{utterance} = \text{form}_i) \propto \sum_{\textit{meaning}} p(\textit{meaning}) * p(\textit{utterance} = \text{form}_i|\textit{meaning})$$

Here we ask whether comprehension processes take into account this generative architecture. Do comprehenders demonstrate a preference for newsworthy informative messages alongside, or instead of, their well-known plausibility preferences? A plausibility preference would manifest as processing ease for a sentence about a high-probability situation (i.e., one of a small number of specific outcomes or scenarios that are predictable in context). An informativity-driven preference would manifest as processing ease with a sentence about a low-probability situation (i.e., any of the many many situation-atypical outcomes or scenarios that do not typically arise). The term *predictability* has many different senses, but it is important to underscore that in our studies, the newsworthy content is not predictable from the situation — it is situation-atypical — nor is it predictable as an upcoming word to complete the sentence — it would be unpredictable in a sentence completion or Cloze task (Taylor, 1953). The content is newsworthy because it is the kind of information that is not easily pre-determined by the comprehender.

Evidence of an informativity-driven bias would stand in contrast to the long history of work linking typicality directly to facilitation (e.g. Hagoort et al., 2004; McRae et al., 1998; Walker, 1975) and would suggest that expectations for atypicality have a role to play in our processing models as well. It would extend prior work on context-driven effects (e.g., Filik & Leuthold, 2008; Nieuwland & Van Berkum, 2006) by showing that comprehenders favor sentences that tell them novel and interesting things, even when the content itself is not specifically guessable from context. Whereas prior work has shown facilitation for content that is situation-typical and/or Cloze-task predictable, our studies are the first to test for the possibility that a sentence containing situation-atypical and Cloze-task-unpredictable content may be overall preferred. Lastly, it would also extend existing work on rational speaker-listener behavior beyond the domain of *how* to communicate a message (i.e., which forms can successfully convey which meanings; Frank & Goodman, 2012) to the choice of *whether* to convey a message at all (i.e., content selection).

Below we review prior work on situation typicality in production (Section 1.1) and comprehension (Section 1.2) and models thereof (Section 1.3). Then we present a series of self-paced reading experiments that use various contextual manipulations to foreground the use of language as a channel across which speakers convey newsworthy and informative messages. In different semantic contexts (unusual protagonists), syntactic constructions (wh- clefts), and communicative environments (text messages), we compare reading times for sentences

[1] Speakers can of course do many things with language, not just deliver news about specific non-inferable situations. They separately or concurrently use language to convey a wide variety of goals, attitudes, identities, etc., but the focus here is on testing whether comprehenders have expectations for newsworthiness since such expectations remain under-explored. We note that expectations are discussed in the literature both in terms of anticipation of upcoming words and in terms of felicity constraints on what is expected of cooperative speakers. Here, we intend the latter, while also assuming that processing is eased when encountering cooperative language. The work here does not attempt to adjudicate between explicit expectation-driven effects and facilitation when integrating preferred linguistic elements (Kuperberg & Jaeger, 2016).



(a) Prior of situation x (b) Production likelihood by prior of situation x

Fig. 1. Hypothetical distributions for situation and utterance predictability, see example (1).

describing typical and atypical situations. The results, primarily in comprehenders' sentence-final reading times, confirm a role for informativity-driven biases during processing: It is possible to observe processing ease for situation-atypical, but pragmatically informative, messages compared with situation-typical, but pragmatically uninformative, messages.

1.1. Informativity in production

Across a range of production studies, speakers are shown to prefer omission of typical or inferable information in favor of atypical aspects of a scenario — speakers opt to include atypical instruments for events (Brown & Dell, 1987; Grigoroglou & Papafragou, 2016; Lockridge & Brennan, 2002: *stab with* [*an icepick* > *a knife*], where [x > y] indicates a preference for x over y), atypical materials/shapes for objects (Mitchell, Reiter, & Van Deemter, 2013: [*wool* > *ceramic*] *bowl*), and atypical colors for foods (Sedivy, 2003: [*pink* > *yellow*] *banana*). When content is included despite being highly predictable from real-world knowledge, it may be needed for referent disambiguation or other goals related to communicative success (e.g., redundant color: Pechman, 1989; Dale & Reiter, 1995; Sedivy, 2003; Arts, Maes, Noordman, & Jansen, 2011; Westerbeek, Koolen, & Maes, 2015; Rubio-Fernández, 2016; Degen, Hawkins, Graf, Kreiss, & Goodman, 2020).

More generally, there is evidence that speakers modulate informativity by adjusting the rate at which they convey information. Low frequency words are more surprising (more informative) than high frequency words, and speakers produce them at a slower rate (Gahl, 2008). Similarly, when the probability of a linguistic element is high, speakers produce more reduced forms (Frank & Jaeger, 2008) or omit optional words (Jaeger, 2010). This prior work measures informativity as the predictability of a word or syntactic structure given a preceding window of context, but in theory the same principles should apply to the predictability of meaning. A message about a highly predictable situation ought to be reducible to an imperceptible surface form $(\texttt{contains}\ (\texttt{room}, \texttt{air}) \rightarrow \emptyset)$ because if uttered, it would represent such a trough in the communication channel's transmission rate.

Speakers' preference for the omission of inferable content is in keeping with long-standing notions of cooperativity: A cooperative interlocutor is expected to make a contribution that is as informative as is required but not more informative than necessary (Grice, 1975; Sperber & Wilson, 1995). When this constraint is violated, pragmatic inferences arise. For example, a speaker's inclusion of a highly predictable color adjective (*yellow banana*) may signal that the speaker's goal is to disambiguate the object from another item of the same category (e.g., a brown banana). Indeed, Sedivy (2003) shows that mentioning a highly predictable color (*pick up the yellow banana*) helps a listener eliminate a color competitor (a yellow notebook) more quickly in a scene that contains a pair of color-contrasting objects of the same category (yellow/brown bananas) than one that doesn't (yellow banana with no category competitor). Comprehenders thus draw inferences from how much information a speaker has chosen to include and they do so in targeted context-specific ways.

This approach to studying informativity — via choice of referring expression — has been the focus of most work on speakers' over- and under-informativity, sometimes termed 'rational redundancy' (Degen, Hawkins, Graf, Kreiss, & Goodman, 2020; for reviews, see Krahmer & Deemter, 2012; Davies & Arnold, 2019). This focus has the side effect of targeting sentences that are already underway, i.e., contexts in which a speaker is already committed to describing something and needs to make decisions about the inclusion/omission of particular modifiers. An open question then is how speakers decide which propositions are worth uttering in the first place — i.e., is the observation *This banana is yellow* newsworthy enough to merit articulation out of the blue? Does a comprehender expect to hear a sentence like that?

One domain that considers the status of speakers' out-of-the-blue sentences is the domain of negation words like *not*. Sentences with negation have received attention because they can be shown to violate a generalisation that true statements are easier to evaluate than false statements (Fischler, Bloom, Childers, Roucos, & Perry, 1983). However, processing difficulty for negated (but true) sentences like *A robin is not a tree* may stem not from difficulty in computing the negated meaning but from its information status — there are many things that a robin is *not*, so a comprehender may struggle with this perfectly true sentence because it "violates the default assumptions that people have about speakers communicating rationally and efficiently" (Nieuwland & Kuperberg, 2008, p.1214; see also Tian, Breheny, & Ferguson, 2010 and Nordmeyer & Frank, 2015). Nieuwland and Kuperberg show that when negated content is pragmatically licensed (e.g., *With proper equipment, scuba-diving isn't very dangerous and often good fun*), true negative statements appropriately yield less difficulty than false statements (as indexed by the N400). They conclude that when negation induces processing difficulty, this disruption reflects not only the assessment of sentence meaning (its correspondence with real-world knowledge) but also an assessment of informativity. The studies we present here extend this claim about informativity beyond the domain of negation to the question of whether comprehenders have expectations about speakers' appropriate informativity during language processing more generally.

1.2. Real-world plausibility in comprehension

In contrast to the findings for language production, the general consensus from several decades of research on comprehension is that sentences about plausible situations are favored over those about implausible situations. This effect can be seen in sentence recall (Marks & Miller, 1964: [*Melting snows cause sudden floods* > *Melting parties augur fragrant drivers*]), verification (Walker, 1975: *The height of a home ceiling is* [*9* > *100*] *feet*), target word naming (Stanovich & West, 1979: *the clothes hung inside the* [*closet* > *bridge*]), eye fixations during reading (Morris, 1994: [*the barber trimmed the mustache* > *the person trimmed the mustache*]), and in the N400, a brain response that reflects semantic

processing (Kutas & Hillyard, 1980: *He took a sip from the* [*waterfall* > *transmitter*]). Findings from the N400 show that comprehenders draw on event typicality (Matsuki et al., 2011: *Donna used* [*the shampoo to wash her filthy hair* > *the hose to wash her filthy hair*]), culturally-specific knowledge (Hagoort et al., 2004: *The Dutch trains are* [*yellow* > *white* > *sour*], presented in Dutch to Dutch speakers), and even knowledge of fictional worlds (Troyer & Kutas, 2018: *There are two Beaters on every Quidditch team. Their job is to protect their team from* [*Bludgers* > *Spell-otape*]). Content that is unpredictable from (real-)world knowledge appears to consistently yield disruption in sentence processing.

Such findings are compatible with models in which real-world knowledge is represented in comprehenders' situation models (Zwaan & Radvansky, 1998) and activated automatically (Nieuwland, 2015). Although such knowledge is often encoded in word co-occurrence patterns, plausibility effects appear to reflect a measure of concept coherence beyond the distributional properties of the words themselves (Connell & Keane, 2004), allowing for the dynamic combination of fine-grained situation properties (Bicknell, Elman, Hare, McRae, & Kutas, 2010; McRae & Matsuki, 2009) and extending to a variety of real-world inferences (Fincher-Kiefer, 1996; McKoon & Ratcliff, 1986; Rodríguez-Gómez et al., 2016). Current work considers the challenges of linking comprehenders' event knowledge to their recovery of a speaker's intended message (Kuperberg, 2016), understanding how learners acquire such knowledge to deploy during comprehension (Borovsky, 2017), and building computational models to simulate the use of event knowledge and linguistic knowledge (Elman & McRae, 2017; Venhuizen, Crocker, & Brouwer, 2019).

Many of the findings above are also compatible with models in which comprehenders' sensitivity is driven by lexical semantic knowledge rather than fine-grained world knowledge. Using eye-tracking while reading, Warren, Milburn, Patson, and Dickey (2015) show that when lexical semantic features are carefully controlled, disruption in reading arises only from selectional restriction violations (*Corey's hamster entertained a nearby backpack*, where *backpack* lacks the necessary semantic feature of sentience to satisfy the verb *entertain*), not from violations of real-world knowledge (*Corey's hamster lifted a nearby backpack*, in which *hamster* and *backpack* each individually satisfy the selectional restrictions of the verb *lift* but the resulting situation isn't real-world plausible). Such findings inform debates about modularity and the timing of the availability of lexical knowledge and real-world knowledge during sentence processing: Implausibility associated with selectional restriction violations yields the strongest disruption in real-time processing. Nonetheless the results still link plausibility violations (at least those captured by selectional restrictions) to processing difficulty. This prior work thus still leaves an open question — do comprehenders ever experience relative ease in processing sentences about implausible situations compared to sentences about plausible ones? Evidence of that would point towards a bias in favor of informativity.

Although this prior work never explicitly denies a role for informativity or novelty, taken together it seems to point to a comprehender who expects utterances drawn from the distribution in Fig. 1a, or at least someone who relies on lexical semantic knowledge which itself encodes many aspects of real-world knowledge. Sentences about plausible, predictable situations are easy to understand (in the lab), even if it is hard to imagine a context where it would be appropriate for a speaker to convey such uninformative content in a series of stand-alone observations about the world. In the studies we present here, we test if comprehenders ever favor informative utterances about situation-implausible content, i.e., whether they have a bias for speaker behavior that incorporates the distribution in Fig. 1b.

A reasonable question to ask at this stage would be whether such anticipation has already been established in prior work, given the existence of studies in which plausibility effects appear to be reversible (e.g., Boudewyn, Long, & Swaab, 2015; Filik & Leuthold, 2008; Hald, Steenbeek-Planting, & Hagoort, 2007; Nieuwland & Van Berkum, 2006). Such studies show that comprehenders can use context to adjust what they take to be plausible, i.e., their situation priors are malleable. However, the emphasis is still on the prior itself, as per (3). This is different than the informativity-driven generative architecture in (4).

For example, although common sense dictates that peanuts in the real world are often salted but are rarely in love, comprehenders who read a story setting up a peanut protagonist who sings about a girl he has met show a reversal of the normal ease/difficulty with the words *salted/in love* (as indexed by the N400; Nieuwland & Van Berkum, 2006). They seemingly adapt quickly to the context-driven constraints that govern situation plausibility in this new 'real world'.

However, such a result still recapitulates the outlook that language acts as a transparent mapping from plausible situations to surface utterances. In the peanut fantasy world, a peanut being in love is plausible, and comprehenders are able to adjust their preferences in favor of sentences about a (now plausible) amorous peanut over a (now implausible) salted peanut. The context makes the content *in love* highly probable (*peanut smiling… peanut singing about a girl… peanut dancing… peanut in love*). In that sense, Nieuwland and Van Berkum's results confirm what other comprehension studies have shown — probable meanings yield processing ease relative to improbable meanings. An open question then is whether comprehenders can experience processing ease for content that itself has low probability. The zombies mentioned in sentence (1-b) have low probability, and furthermore, even if a speaker says they see something interesting in the room, the mention of zombies is nonetheless still low probability because there are so many different interesting weird things that could be in a room.

Are there contexts where what is easy to process is not just a differently plausible situation but is actually the implausible? Language does provide some overt cues to herald unexpected outcomes. We have words that signal when the relationship between propositions is one that reverses common-sense reasoning. A connective like *even so* signals a concessive coherence relation, and it can help comprehenders anticipate an outcome opposite to that which typically follows in the normal course of events: *Elizabeth took the test and failed it.* [*Even so, she went home and celebrated wildly* > *She went home and celebrated wildly*] (Xiang & Kuperberg, 2015). Normally, failing a test doesn't lead to celebration, but with *even so*, the word *celebrated* is far easier to integrate (see also Ferguson & Breheny, 2011; Jiang, Li, & Zhou, 2013; Köhne-Fuetterer, Drenhaus, & Delogu, 2020). However, these studies are still restricted to meanings that are derivable in a predictable way via negation: Normally it is the case that ¬(`failure` → `celebration`), so *even so* signals that `failure` → `celebration`. But is comprehension ever facilitated by non-specific surprising content (like the zombies in (1-b))?

Evidence that comprehenders can "expect the unexpected" comes in part from studies on disfluency (e.g., uh/um/er fillers; Corley, MacGregor, & Donaldson, 2007; Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Heller, Arnold, Klein, & Tanenhaus, 2015). The presence of disfluency is shown to help comprehenders anticipate upcoming material that otherwise has low probability, presumably because disfluency may indicate that the speaker is experiencing difficulty in word retrieval or production. Corley et al. measured semantic processing (via the N400) at a typical versus atypical noun (*Everyone's got bad habits and mine is biting my* [*nails* > *tongue*]) and showed that the difficulty normally associated with atypicality is reduced following disfluency (*…and mine is biting my, er, tongue*). Beyond a reduction in difficulty, we ask here whether a sentence containing atypical informative content can ever be outright easier than typical content and where in a sentence such ease is observable.

In a separate strand of research on linguistic cues that signal upcoming surprise, there is also work on plot-driven expectations in narrative text. Even though certain situations or outcomes may be rare (e.g., most cups of coffee in the world remain unspilled), a story plot can build up expectations for one of these rare outcomes (*The cup of coffee was balanced on the arm of the chair. Suddenly, Richard sneezed….*; Grimes-Maguire & Keane, 2005). The insight from such work is that

comprehenders are depicted as tracking knowledge of story structure separately from knowledge of the causal structure of the world (see also Rapp & Gerrig, 2002). The ability to do this depends crucially on a distinction between the world and what people say about the world. The ability to track those distinct probabilities need not be limited to plot-driven twists in specific genres, but should be observable more generally.

1.3. Modelling informativity expectations

In its simplest form, comprehenders' expectation for informativity should be based on their observation of what speakers actually say. In this sense, any model that incorporates frequency statistics ought to capture the kind of speaker behavior reviewed in Section 1.1. Utterance frequencies would correctly capture the intuition that *the room is full of air* is dispreferred compared to *the room is full of people*, but they don't offer an account for why speakers' productions contain the relative proportions they do.[2] The probability of seeing a word is different from the probability of the underlying meaning, as Taylor (1953) acknowledged in discussing his newly established Cloze measure, which "counts instances of language-usage correspondence rather than meanings themselves" (p.417–418). What is needed is a model that incorporates situation predictability and links situation (un)predictability to production likelihood to understand the kinds of utterances comprehenders tend to favor.

The Rational Speech Act (RSA) model has been put forward as a framework that encodes measures of utterance utility and cost. It models speaker decisions in light of how a savvy listener would draw inferences from different utterances (Frank & Goodman, 2012; see also Lewis, 1969 or Benz et al., 2005 on Game Theory, or Franke & Jäger, 2016 for a general introduction to probabilistic pragmatic approaches). What is being put forward here in this paper as the necessary components of a model of informativity is consistent with the underlying principles of RSA.

That said, to our knowledge, RSA has been applied exclusively to communicative scenarios in which the decision to speak has already been made: A situation or referent must be described, and RSA accurately predicts interlocutor choices in such situations. For example, in a context with multiple objects (blue square, green square, and blue circle) and a limited inventory of words to use (*blue, circle*), RSA captures why a speaker can successfully refer to the blue square with *blue* even in the presence of the blue circle, because the interlocutors can reason that the expression *circle* is available to unambiguously refer to the competing blue circle (see also Degen & Franke, 2012; Rohde, Seyfarth, Clark, Jaeger, & Kaufmann, 2012; Stiller, Goodman, & Frank, 2015). Likewise, in a context in which a speaker has been asked to evaluate something (e.g. Ann baked a cake and asked Bob about it), RSA predicts how the speaker's evaluation (*it was okay*) is interpreted against the backdrop of the utility of providing new and accurate information and the utility of being polite (Yoon, Tessler, Goodman, & Frank, 2016; see also Bennett & Goodman, 2018). In these cases, the speaker must speak, and RSA captures the use and interpretation of particular forms by appealing to concepts of message utility and cost and recursive speaker-listener estimates of how language will be used.

This machinery is precisely what is needed to account for speakers' predictions of what a comprehender would do with a given utterance (and what a comprehender can estimate that a speaker is predicting will be inferred). In the case of *the room is full of air*, the triviality of this observation (its limited utility) is likely not worth the cost of uttering it. The state of the world and how we talk about it are thus inextricably linked. RSA has been successfully applied to model the way comprehenders use a speaker's utterance form to update their understanding of the world (e.g., Degen, Tessler, & Goodman, 2015; Yoon, Tessler, Goodman, & Frank, 2016). Working with scalar implicatures, Degen et al. show how certain semantically weak utterances can induce changes in comprehenders' estimates of real-world probabilities. The utterance *some of the marbles sank* implicates that not all the marbles sank, in a scenario where a set of marbles were thrown into a pool. Since this outcome is at odds with our real-world knowledge, the comprehender is driven to update their judgments about the world and to make allowances for a 'wonky' world. In the studies we present here, we measure comprehenders' response both to sentences about highly wonky outcomes (ones that would require comprehenders to update their prior understanding of the world) and to perfectly un-wonky situations (ones that should be inferable). As we will show, it is not the case that the latter is always easier.

In testing and modelling informativity-driven effects, a question that arises is about timing — at what point in an utterance can a comprehender assess the informativity of the message? In the studies we report here, a target word is used to evoke a typical versus atypical situation. Any reading time differences (at or after the target word) will depend both on the availability and speed of comprehenders' informativity-driven computations and also on the degree to which a given utterance signals to the comprehender the position of the informative content. Most sentences do not specify a single location where the informative content must appear, meaning that comprehenders don't need to pin all their hopes on the target word being the sole locus of informativity. It is often only at the last word of the sentence or the last sentence in a speaker's turn that message (un)informativity becomes unambiguous, in which case effects might emerge late.

Late-emerging effects would be in keeping with clausal-integration accounts that have been posited for other pragmatic phenomena (Garnham, Traxler, Oakhill, & Gernsabacher, 1996; Stewart, Pickering, & Sanford, 2000). Alternatively, if the utterance makes clear which element is intended to be informative, effects might emerge at the target word or shortly after, in keeping with fast pragmatic processing in domains like reference, implicature, social context, and coherence (Grodner, Klein, Carbary, & Tanenhaus, 2010; Hanna, Tanenhaus, & Trueswell, 2003; Rohde, Levy, & Kehler, 2011; Van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2008). Note that the approach taken here does not rule out the presence of plausibility-driven effects (see formula (4), which contains a term for the situation prior as well), but it does predict that at some point in the sentence the ease with situation-plausible content can be eliminated or reversed.

2. Experiment 1: Unusual protagonist as a cue to unpredictability

This first experiment uses self-paced reading to test whether the way a protagonist is portrayed can alter the processing cost normally associated with action-atypical outcomes. We portray a protagonist either as a boring person, who always does things the way one would expect, or as a surprising person, who never does things the way one would expect. The protagonist is then described performing an action with either a typical or an atypical instrument, as in (5).

[2] A search of the Google ngram corpus (Brants & Franz, 2006) confirms that speakers do talk more about rooms that contain people than about rooms that contain air. Given the counts below, comprehenders would not encounter many utterances about very rare situations (rooms full of zombies) nor very common situations (rooms full of chairs or air). Nonetheless, a comprehension bias in favor of informativity would distinguish these two types, predicting the possibility of relative ease for processing the rare and interesting over the very mundane.

Case-insensitive trigram	Count
room(s) full of people	29,446
room(s) full of diamonds	1320
room(s) full of chairs	144
room(s) full of zombies	116
room(s) full of air	94

(5)

a. In order to chop some carrots, he was using a knife. [action-typical]
b. In order to chop some carrots, he was using a shovel. [action-atypical]

Two aspects of these materials are designed to heighten a bias in favor of informativity. First, the portrayal of a protagonist as surprising is intended to increase comprehenders' expectation that the speaker may have something newsworthy to say. This is different from context-specific or speaker-specific biases that help a comprehender anticipate a high-probability situation-plausible outcome (e.g., a fantasy world that dictates that a dancing peanut is likely to be *in love* rather than *salted*; Nieuwland & Van Berkum, 2006; a social stereotype that dictates that a male speaker is more likely to say 'I dropped my aftershave on the floor' rather than 'If only I looked like Britney Spears'; Van Berkum et al., 2008). Here, a comprehender cannot anticipate what specific instrument a surprising protagonist would use — there are many many atypical instruments that are possible, each one of which thus has low probability but any one of which would be satisfyingly informative for a cooperative speaker to mention. Second, we chose to target instruments because of their status as optional. A speaker's choice to include an optional element may heighten comprehenders' expectation for informative content (see Brown & Dell, 1987).

The action-typical instruments are ones that are largely inferrable from the action itself: Carrot chopping events have high probability of involving a knife. The action-atypical instruments are not inferrable from the action and are instead high-probability instruments for other items. In a model in which real-world typicality maps transparently to utterance production, atypical content should be surprising: (5-b) should yield more comprehension difficulty than (5-a). However, according to the model put forward here that combines situation typicality with informativity, the rarity of an atypical situation makes that situation more interesting and appropriate to talk about, particularly when the discourse emphasizes upcoming surprisal. We thus predict an interaction: (5-b) should be harder than (5-a) in the context with the normal protagonist but less so for the context with the surprising protagonist.

A challenge in measuring responses to atypicality is that repeated exposure to such messages may undermine participants' sensitivity to what is appropriately newsworthy. Over the course of a long experiment, participants may abandon their expectations for newsworthiness or shift their sense of what counts as typical. Since the model we're testing is specifically about comprehenders' expectations regarding the things a normal and cooperative speaker would choose to say, reliable effects may depend on maintaining participants' typical discourse expectations. We therefore take advantage of one of the benefits of crowdsourced web experiments — that many participants can be recruited, each of whom only sees a small number of items. For the studies reported in this paper, each participant saw only one item per condition, interleaved with fillers. A large number of participants allows us to approximate the number of observations from in-lab psycholinguistic studies. For example, a typical in-lab psycholinguistic study with a 2×2 design might recruit ~24 participants, each of whom would see ~24 items, yielding a dataset with 576 datapoints. Here we aim for a similar-size dataset by recruiting over 100 participants, each of whom only see one item in each of 4 conditions.

2.1. Methods

2.1.1. Participants

One hundred thirty-six participants were recruited via Amazon Mechanical Turk and paid $0.60.

2.1.2. Materials

Experimental items consisted of short passages in a 2×2 design that varied the portrayal of a protagonist in the first sentence (boring vs. surprising) and the typicality of an instrument in the second sentence (action-typical vs. action-atypical). A sample item is shown in (6).

(6)

a. [boring protagonist / action-typical instrument]
My cousin Mary is a boring person who always does things the way you'd expect. For instance, in order to dig a hole, she was using a shovel yesterday in the afternoon.
b. [boring protagonist / action-atypical instrument]
My cousin Mary is a boring person who always does things the way you'd expect. For instance, in order to chop some carrots, she was using a shovel yesterday in the afternoon.
c. [surprising protagonist / action-typical instrument]
My cousin Mary is a surprising person who never does things the way you'd expect. For instance, in order to dig a hole, she was using a shovel yesterday in the afternoon.
d. [surprising protagonist / action-atypical instrument]
My cousin Mary is a surprising person who never does things the way you'd expect. For instance, in order to chop some carrots, she was using a shovel yesterday in the afternoon.

The instrument *shovel* is the critical region in (6). In all items the critical region was always followed by a 4-word spillover, which concluded the sentence. We selected actions which strongly favor particular conventional instruments. The materials were counterbalanced so that, across participants, each action appeared with an action-typical instrument and an action-atypical instrument and each instrument appeared as the action-typical and action-atypical instrument for two different actions. The set of 13 action/instrument combinations are shown in Table 1. Filler items were similar in that they introduced a referent in the first sentence (a person or a place) and described a property of that referent, which was then illustrated with an example in the second sentence (see Appendix A).

To quantify comprehenders' evaluations of these items and their expectations for upcoming content, we conducted a set of offline norming studies. The participants in these studies were all monolingual English speakers recruited from Amazon Mturk; none took part in the reading time study.

Regarding the expectedness of the target word in our experimental items, we collected two measures. The first addresses situation possibility/impossibility, given known differences in the strength, localization, and context sensitivity of processing disruption for implausible versus impossible events (Warren, McConnell, & Rayner, 2008). Participants ($N = 36$) gave yes/no responses to questions such as *Is it possible for someone to chop carrots with a shovel?* The results confirmed that situations with typical instruments in our materials were judged more possible (e.g., *dig with a shovel*, mean possibility 0.99) than those with atypical instruments (e.g., *chop with a shovel*, mean possibility 0.34; logistic regression: $\widehat{\beta}$=12.21, SE = 3.02, z = 4.05, p<0.001), and that most

Table 1
Experiment 1 action/instrument combinations for target sentences

Action	Predictable Instrument	Unpredictable Instrument
dig hole	shovel	fork
chop carrots	knife	shovel
brush teeth	toothbrush	knife
clean porch	broom	toothbrush
repair brakes	wrench	broom
secure yacht	rope	wrench
accessorize dress	belt	rope
transport groceries	cart	belt
drain spaghetti	strainer	cart
wrap present	ribbon	strainer
wash dishes	sponge	ribbon
write letter	pen	sponge
eat steak	fork	pen

atypical instruments were judged as possible at a rate greater than zero. However, there were two items that had a mean possibility score of 0.0 in the atypical condition (*Is it possible for someone to write a letter with a sponge*, *Is it possible for someone to wrap a present with a strainer*), receiving only 'no' responses, similar to our seven strongly impossible fillers (e.g., *Is it possible for someone to blackmail spaghetti?*, mean possibility 0.01). In the Results, we report analyses with and without those two items as well as an analysis using the continuous measure of instrument possibility in place of the binary typical/ atypical factor.

For the second measure, participants ($N = 33$) saw the experimental items up to but not including the target word, similar to a Cloze task. We varied the type of protagonist and elicited an instrument from the participant (… *she was using a/an* ___). For each item, we computed the entropy of the set of completions in the boring and surprising protagonist conditions separately. The results showed that the surprising protagonist condition yielded completions with higher entropy (2.56 bits) than the boring protagonist condition (1.62). This difference between means (0.94) is significant at $p<0.001$ by a permutation test. For the permutations, we randomly shuffled the surprising/boring condition labels for the completions for each item and then computed the resulting difference in entropy scores. The largest difference achieved was 0.46 across 100,000 permutations. This suggests that hearing about a surprising protagonist does not simply adjust comprehenders' expectations in favor of a different specific outcome (akin to the adjustment in priors seen in experiments that use fantasy worlds to make specific real-world-implausible outcomes plausible); rather it changes comprehenders' expectations more generally to include more variable outcomes. The probability of the actual action-atypical instruments in our materials was close to zero (*chop carrots with a shovel*). Given this, any ease we observe in the action-atypical condition can't be attributed to a high Cloze probability.

Lastly, we conducted a third norming study to test the extent to which informativity expectations are localized to the target word or whether comprehenders posit that subsequent words can provide novel and informative content. Participants ($N = 34$) read the experimental items, up to and including the target word but with no sentence-final punctuation. We varied the type of protagonist, and all items appeared with the typical instrument. Participants were instructed to complete the sentence, either with a full stop '.' or with additional words. For filler catch trials, their completions confirmed that they appropriately inserted a '.' when the text contained a known phrase that was already complete (e.g., *great minds think alike __*, *you can't teach an old dog new tricks __*) or added words when something was missing (e.g., *better late than __*, *he likes to sing Jingle __*). The results on the target sentences showed that the surprising-protagonist condition yielded a higher rate of non-full-stop completions (0.95) than the boring-protagonist condition (0.55, logistic regression: $\widehat{\beta}$=3.20, SE = 1.18, z = 2.71, $p<0.01$). Within the non-full-stop completions, the surprising condition yielded longer completions (number of characters for boring vs surprising: 15.79 vs 22.28; linear regression $\widehat{\beta}$=6.74, SE = 1.84, t = 3.67, $p<0.001$) and more modifiers (rate of modification for boring vs surprising: 0.07 vs 0.30; logistic regression $\widehat{\beta}$=1.96, SE = 0.72, z = 2.70, $p<0.01$). Given that participants appear able to envision sentence comepletions that provide informative content after the target noun, informativity effects may not be localized to the target word.[3]

2.1.3. Procedure

Experiment 1 uses a web-based self-paced reading task. Having indicated their consent, each participant saw several practice items, followed by the main experiment. Each participant saw only 4 target items. The small number ensures that participants only encounter two discourse-infelicitous sentences. A participant's 4 critical items were selected from the 13 possible items in Table 1. Each participant also saw 2 filler items, selected from a set of 10 possible fillers. The task was presented in a web browser using IbexFarm's moving-window self-paced reading paradigm (Drummond, 2013). Sentences initially appeared as a series of dashes obscuring the words (– – – –), and participants pressed the space bar to reveal each region. The presentation was non-cumulative so previous regions were replaced with dashes when the next region appeared. Regions were revealed one word at a time. After each item, participants saw a comprehension question about either the first half or second half of the item to encourage them to read the items in full; they answered the question by clicking one of two possible responses. We recorded reading times for each region as well as the participants' responses to comprehension questions. Counterbalancing was achieved through IbexFarm's automated Latin Square counter, assigning each new participant to the next experimental list, which in our case additionally required specification of lists of only 4 target items.

2.1.4. Analysis

The analysis of raw reading times was conducted with linear mixed-effect regression models (LMER; Baayen, Davidson, & Bates, 2008), using the *lme4* package in R (Bates, Mächler, Bolker, & Walker, 2015; R Core Team, 2017). A standard approach for self-paced-reading studies involves conducting a series of separate analyses, one for the target region and one for each spillover region. However, the non-independence of reading times at different positions in the same sentence raises a concern about multiple comparisons. Therefore, we have opted to report two modelling strategies for all experiments in this paper. In the main text, we present the classic region-by-region analysis, and we apply Bonferroni corrections for multiple comparisons. In the Appendix, we present an alternative analysis in which we start by building a single large model containing the manipulated factors along with Region, and we only conduct follow-up analyses for interactions that reach significance in this large model. This latter approach limits the number of follow-up analyses by only targeting interactions that reached significance in the omnibus analysis, and so no Bonferroni correction is needed. This follows recent work on multi-window analyses in other psycholinguistic methods (see Grüter, Takeda, Rohde, and Schafer (2018) for a similar analysis of eye-tracking data).

All fixed effects are centered in order to facilitate model interpretation. Protagonist has two levels (boring vs. surprising, coded as −0.5 and 0.5). Instrument Typicality has two levels (action-typical vs. action-atypical, coded as −0.5 and 0.5). Region, which is included in the single-model approach reported in the appendix, has five levels (target word as reference level).

For this experiment, we include by-instrument and by-participant random intercepts and slopes, with the exception of the by-participant random interaction term. The exclusion of this term reflects the fact that each participant provided only one datapoint per condition, meaning the model can estimate variance for both levels of the Typicality factor (two action-typical datapoints, two action-atypical datapoints) and for both levels of the Protagonist factor (two boring datapoints, two surprising datapoints), but not the interaction. For models that fail to converge with maximal random effect structure, we first remove correlations between random intercepts and random slopes and then iteratively remove low-variance random slopes. Significance is determined using Satterthwaite's method of approximating degrees of freedom in the lmerTest package (Kuznetsov, Brockhoff, & Christensen, 2017), with Bonferroni corrections applied to the model interpretation in the main text.

In addition, we use footnotes to report two variants of the region-by-region analysis: In one variant, we replace the binary factor Typicality with a continuous measure from the possibility/impossibility norming study; in the other variant, we exclude the two items that received

[3] We thank an anonymous reviewer who recommended the Cloze task and two other anonymous reviewers whose questions pointed to the relevance of the (im)possibility and localization measures.

possibility scores of 0.0 in the possibility/impossibility norming study. Across the additional footnote and appendix analyses for this experiment, all critical findings for informativity-driven effects remain significant.

2.2. Results

After excluding participants who did not list English as their native language and those with less than 80% accuracy on comprehension questions, our dataset consists of responses from 110 participants. Every item/condition combination was seen by at least 5 participants and an average of 8.5 participants. We removed trials with consecutive reading times under 50 ms, indicating that the participant was simply holding down a key. The analysis considers all non-outlier items, regardless of comprehension-question accuracy. As a first cut to remove outliers, we excluded reading times below 100 ms and above 5000 ms (24 datapoints in the critical and spillover regions). We then removed reading times that were more than three standard deviations away from the mean, per region and per condition (a further 13 datapoints in the regions of interest).

Table 2 shows the raw reading times by condition for the critical region and the four spillover regions (visualized in Fig. 2). Visual inspection indicates separation of the reading times by condition in the sentence-final region (Spill4), and the observed pattern is in keeping with the predicted Protagonist × Typicality interaction: action-atypical instruments are read faster in a passage about a surprising protagonist than in a passage about a boring protagonist. Notably the fastest reading time in that region is that of the action-atypical instrument with the surprising protagonist.

Table 3 shows the model results for each region. After Bonferroni correction (adjusted threshold for significance =0.01), the only significant effect is the crucial Protagonist × Typicality interaction at the sentence final region ($\widehat{\beta}$= −277.18, p<0.001). Follow-up analysis confirms that with the surprising protagonist, there is a main effect of typicality whereby unpredictable instruments yield faster RTs than predictable instruments ($\widehat{\beta}$= −161.66, SE = 48.27, $t = -3.35$, $p = 0.01$). In contrast, with the boring protagonist, the difference is numerically reversed but there is no significant effect of predictability ($\widehat{\beta}$= 51.95, SE = 37.36, $t = 1.39$, $p = 0.17$).[4]

See Appendix B for the alternative single-model approach, which captures the same pattern of results by building a single larger model that contains Region as an additional fixed effect. The results show the critical interaction between Protagonist, Typicality, and Region (at Spill4), confirming that the interaction at the sentence-final region differs from that at the target region, the reference level for Region.

2.3. Discussion

Experiment 1 was designed to test the hypothesis that the well-known difficulty associated with real-world implausibility can be altered if a comprehender is encouraged to expect interesting and newsworthy messages. The results are consistent with a comprehension bias in favor of informativity whereby difficulty varies with discourse context, i.e., an interaction between Protagonist and Instrument Typicality (significant in the primary analysis reported in the text, in the two alternative analyses in footnote 4, and in the single-model approach reported in Appendix B). In passages about a surprising protagonist, the analysis shows a reversal of the standard plausibility effect, such that action-atypical instruments yield faster reading times than action-typical instruments. In passages about a boring protagonist, action-typical instruments yield numerically faster reading times. This is in keeping with the claim that comprehenders favor discourse-appropriate messages about interesting atypical situations.

Although prior studies have manipulated the context in order to modulate comprehenders' response to real-world typical/atypical content, such studies use context to set up a new 'real world' in which specific alternative outcomes are plausible or they use discourse markers to signal that typical reasoning should be reversed. In contrast, our portrayal of a surprising protagonist can be said to create expectations for outcomes drawn from a larger set of unusual outcomes (as shown in the second norming study on entropy). There is no favored high-Cloze-probability word, but sentences are favored if they contain some content that is unusual. The work that is most directly comparable is that on disfluency and processing low-Cloze-probability words. Although the presence of *er* had been found to eliminate the difficulty with subsequent situation-atypical content (Corley et al., 2007), our result in Experiment 1 shows that a sentence about atypical content can actually be easier to process than one about typical content.

Regarding the question of where in the sentence we would see informativity-driven effects, the pattern only emerged sentence finally. This either could indicate a general delay in pragmatic processing or it could reflect the difficulty comprehenders face in determining what specific part of a sentence is intended to be informative. The first explanation is in keeping with clausal-integration accounts that posit that comprehenders' consideration of pragmatic felicity only emerges at later stages of processing when the proposition is integrated into the larger discourse context (e.g., Garnham et al., 1996; Stewart et al., 2000). Such accounts have been largely superseded by findings that show fast pragmatic processing (e.g., Grodner et al., 2010; Hanna et al., 2003; Rohde et al., 2011; Van Berkum et al., 2008). Under the second explanation, we can think of our participants as showing difficulty at the point in the sentence where the infelicity is no longer rescuable. An action-typical instrument may fail to yield an immediate slowdown because the comprehender can trust that the newsworthy part of the message could still be coming (*he dug a hole with a shovel… and then ate the shovel*). In this way, our results show compatibility with accounts in which pragmatic processing can operate quickly, and the sentence-final timing simply reflects the late point at which the felicity violation reveals itself.

In order to see where informative content might be expected to appear in our materials, we reviewed the completions that participants wrote in the second norming study on entropy. Although many completions provided informative content at the first word (e.g., *wash dishes using a/an… cat*), others only delivered the informative content in later words (e.g., via an informative head noun after an indeterminate pre-nominal adjective: *repair the brakes using a/an… small jackhammer to undo the nuts*, via modification after the action-typical head noun: *brush teeth using a/an… toothbrush for a person's finger tips to use*, or via a more complex noun phrase: *secure a yacht using a/an… piece of gum*). In the latter cases, the first word does little to convey informative content. If comprehenders are aware of this variability, they need not expect the target position in our materials to be the only place that informative content could appear. Similarly, the third norming study showed that participants were very willing to add informative material after the action-typical target noun, particularly in the surprising protagonist condition (e.g., *brush teeth using a toothbrush… dipped in applesauce*). It therefore may only be at the final region of an uninformative sentence, when nothing appropriately newsworthy has been found, that the sentence becomes infelicitous. Experiment 2 attempts to make clearer to the comprehender the position of the newsworthy content by using a syntactic construction that highlights the position of new information in a particular constituent.

[4] Across the two additional region-by-region analyses, the pattern of results stays exactly the same with the same Bonferroni-corrected threshold for significance. Replacing the binary factor Typicality with a continuous measure from the possibility/impossibility norming study yields only one significant effect, the sentence-final interaction at spillover 4. Likewise, eliminating data from the two items that received possibility scores of 0.0 in the norming study (*sponge* and *strainer*) again yields only the sentence-final interaction.

Table 2
Experiment 1 reading times (ms) by condition and by region (participant means ± standard error)

	Instrument	Spill1	Spill2	Spill3	Spill4
Boring Action-typical	455.73±16.79	457.53±17.24	439.95±15.51	385.72±9.69	657.12±37.41
Boring Action-atypical	454.01±18.54	512.58±25.50	435.34±14.42	398.74±13.06	717.82±49.09
Surprising Action-typical	429.26±15.00	463.52±16.69	419.67±11.71	391.04±12.82	747.90±49.43
Surprising Action-atypical	425.37±15.03	475.60±18.90	414.80±13.29	381.27±13.62	547.63±29.33

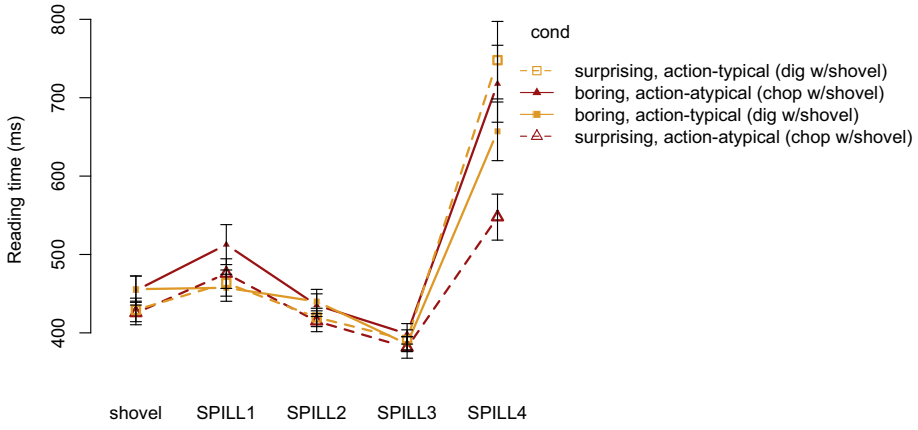


Fig. 2. Experiment 1 reading times (ms) by condition, from target instrument to end of sentence.

Table 3
Results of linear mixed-effect models for Experiment 1 reading time data. Boldface indicates significance after Bonferroni correction for the five regions' non-independent analyses (adjusted threshold =0.01)

Instrument	$\widehat{\beta}$	SE	t	p	**Spillover 3**	$\widehat{\beta}$	SE	t	p
(Intercept)	**442.41**	**15.79**	**28.02**	**<0.001**	**(Intercept)**	**389.52**	**10.70**	**36.41**	**<0.001**
Protagonist	−27.34	12.49	−2.19	0.05	Protagonist	−8.09	11.14	−0.73	0.49
Typicality	3.60	13.63	0.26	0.79	Typicality	0.60	9.40	0.06	0.95
Protag × Typ	−4.88	26.25	−0.19	0.86	Protag × Typ	−20.50	16.75	−1.22	0.22
Spillover 1	$\widehat{\beta}$	SE	t	p	**Spillover 4**	$\widehat{\beta}$	SE	t	p
(Intercept)	**478.01**	**15.84**	**30.17**	**<0.001**	**(Intercept)**	**675.08**	**39.37**	**17.15**	**<0.001**
Protagonist	−15.13	14.71	−1.03	0.31	Protagonist	−42.29	41.30	−1.02	0.33
Typicality	35.82	15.28	2.35	0.04	Typicality	−70.33	33.95	−2.07	0.04
Protag × Typ	−48.72	33.56	−1.45	0.20	**Protag × Typ**	**−277.18**	**60.81**	**−4.56**	**<0.001**
Spillover 2	$\widehat{\beta}$	SE	t	p					
(Intercept)	**425.60**	**12.67**	**33.61**	**<0.001**					
Protagonist	−20.64	9.75	−2.12	0.05					
Typicality	−3.97	9.16	−0.43	0.67					
Protag × Typ	1.13	24.07	0.05	0.96					

3. Experiment 2: Wh- cleft as a cue to unpredictability

The results of Experiment 1 are consistent with a comprehension bias in favor of informativity, but the expected interaction didn't show up until the final region of the sentence. Since the late emergence of the interaction could reflect participants' uncertainty about which words constituted the newsworthy part of the sentence, Experiment 2 manipulates the syntax of the target sentence to test whether signaling where newsworthy material will appear can induce earlier informativity-driven effects. The manipulation involves a contrast between a canonical syntactic structure and a non-canonical structure that places focus on one constituent, as in (7).

(7)

a. Experiment 2 uses a wh- cleft. [canonical, no cleft]

b. What Experiment 2 uses is a wh- cleft. [non-canonical, wh-cleft]

The wh- cleft structure represents one of many linguistic devices available to speakers for signaling focus, alongside devices such as intonation, focus particles, and prior discourse context (see Lowder & Gordon, 2015; Rooth, 1992). Wh- clefts assign focus to the post-copular material (the noun phrase *a wh- cleft* in (7-b)). To be used felicitously, a wh- cleft requires there to be a salient proposition in the context that has been left 'open' or unspecified (Birner, 2004; Birner & Ward, 2009; Prince, 1978). The clefted material fills in the missing material and hence represents focused, often new, information.

Prior work shows that focused material typically requires increased processing time: The same words are read more slowly when they appear as a clefted noun phrase than in a sentence with canonical structure (e.g., *memo* in *What the secretary typed was the official memo*

versus *Yesterday the secretary typed the official memo*; Lowder & Gordon, 2015). This extra encoding time in turn yields better memory for the focused content and easier reactivation at a subsequent anaphor (Almor, 1999; Cowles & Garnham, 2005). What hasn't been tested is whether the focusing effect of a wh- cleft makes it easier to read particular kinds of content in that position — namely new or surprising content.

For our purposes, the wh- cleft signals that an upcoming constituent will contain new information and it specifies which constituent that must be. If expecting new information makes it easier to process words that describe atypical situations, the wh- cleft should make it easier to process situation-atypical content. Canonical syntactic structures, on the other hand, are used broadly across contexts and do not impose strong information-structural constraints about the location of new information. We thus predict an interaction, now as early as the post-copular noun phrase, between syntax and the situation typicality of the entity mentioned in that post-copular constituent. Given that non-canonical structures are more rare and given that prior work shows longer reading times for clefted words, we may see an overall processing slowdown for the wh- cleft compared to the canonical structure. Finally, given prior work on real-world plausibility, we may see a time window in which participants show an overall processing slowdown for atypical content compared to typical content. Again, we recruited a large number of participants, each of whom saw only one item per condition.

3.1. Methods

3.1.1. Participants

One hundred thirty-six participants were recruited via Amazon Mechanical Turk and paid $0.70.

3.1.2. Materials

The materials were adapted from Experiment 1. The same 13 action/instrument combinations were used. The first sentence introduced a protagonist and an action. The second sentence varied the syntactic structure (canonical vs. cleft) and the typicality of the instrument (action-typical vs. action-atypical). The critical region (*shovel* in (8)) was always followed by a 5-word spillover, which concluded the sentence.

(8)

a. [canonical / action-typical instrument]
My cousin Mary was digging a hole yesterday in the afternoon. She was digging the hole with a shovel, and she got really tired.

b. [canonical / action-atypical instrument]
My cousin Mary was chopping some carrots yesterday in the afternoon. She was chopping the carrots with a shovel, and she got really tired.

c. [cleft / action-typical instrument]
My cousin Mary was digging a hole yesterday in the afternoon. What she was digging the hole with was a shovel, and she got really tired.

d. [cleft / action-atypical instrument]
My cousin Mary was chopping some carrots yesterday in the afternoon. What she was chopping the carrots with was a shovel, and she got really tired.

3.1.3. Procedure

The procedure followed that of Experiment 1.

3.1.4. Analysis

The region-by-region analysis followed that of Experiment 1. The same linear mixed-effects modelling approach was used, now with syntax as a fixed effect instead of protagonist (canonical vs. cleft, coded as −0.5 and 0.5). Again, we use a Bonferroni-corrected threshold for determining significance for both the primary analysis and the footnote analyses (i.e., the models with the continuous typicality measure and with the two low-possibility items excluded). Appendix C reports a single-model approach that parallels that reported for Experiment 1. In this case, the different approaches yield different results. The predicted interaction is significant sentence-finally in the region-by-region analysis, but not in the analyses with the continuous typicality measure or with the two low-possibility items excluded. In the single-model approach, the interaction is significant much earlier, directly at the target word.

3.2. Results

We applied the same participant exclusion criteria as in Experiment 1, yielding a dataset with reading times from 89 participants in which every item/condition combination was seen by at least 4 participants and an average of 6.8 participants. As before, the first-cut outlier removal eliminated reading times below 100 ms and above 5000 ms (5 datapoints in the critical and spillover regions). We then removed reading times more than three standard deviations away from the mean, per region and per condition (a further 27 datapoints in the regions of interest).

Table 4 shows the raw reading times by condition for the critical region and the five spillover regions (see Fig. 3). Visual inspection indicates an overall preference for the action-typical instruments, particularly in the first two regions. The means also show that at the target word and at the sentence-final regions, the pattern of results reflects a Syntax × Typicality interaction: In the canonical condition, participants show processing difficulty with action-typical instruments, whereas in the cleft condition (which overall has longer reading times), typical/atypical instruments yield similar reading times.

Table 5 shows the model results for each region. After Bonferroni correction (adjusted threshold for significance =0.008), we see the following effects. At the first spillover region, there is a main effect of Typicality ($\widehat{\beta}$=63.17, p<0.001), whereby action-atypical instruments yield longer reading times. It is only at the sentence-final region that we find the predicted interaction whereby the difficulty associated with action-atypical instruments is reduced in the cleft structure ($\widehat{\beta}$= −64.84, $p = 0.008$). Follow-up analyses of this interaction at the sentence-final region show that the slowdown with action-atypical instruments in the canonical syntax disappears in the cleft syntax. In the canonical condition, there is a main effect of Typicality ($\widehat{\beta}$=39.86, SE = 14.89, $t = 2.68$, $p = 0.01$), whereas in the cleft condition, the effect is numerically reversed and is no longer significant ($\widehat{\beta}$= −5.42, SE = 14.43, $t = -0.38$, $p = 0.71$). Table 5 contains other findings which include an effect of Typicality at the target region (showing slower reading times for atypical instruments), effects of Syntax at the target and spillover1 regions (showing slower reading times for the cleft condition), and the predicted Syntax × Typicality interaction, but those do not survive Bonferroni correction.[5]

[5] The pattern of results varies somewhat across the two additional analysis approaches. Again we use $p = 0.008$ as the Bonferroni-corrected threshold for significance. The variations are as follows. First, when we replace the binary factor Typicality with the continuous possibility measure, the main effect of Typicality at the target region and the main effect of Syntax at Spill1 reach significance: Atypical instruments yield longer reading times than typical instruments at the target word ($\widehat{\beta}$= − 30.45, SE = 10.20, $t = 2.99$, $p = 0.004$), and clefts yield longer reading times than canonical structures ($\widehat{\beta}$=30.39, SE = 9.97, $t = 3.05$, $p = 0.006$). However, the critical interaction at the sentence-final region is no longer significant ($\widehat{\beta}$=33.14, SE = 17.00, $t = 1.95$, $p = 0.10$). Secondly, eliminating data from the two items that received possibility scores of 0.0 allows the main effect of Syntax at the target region to reach significance ($\widehat{\beta}$=52.08, SE = 19.08, $t = 2.73$, $p = 0.008$). However, in that analysis, the critical interaction does not reach significance under the Bonferroni-corrected threshold at either Spillover 4 ($\widehat{\beta}$=41.48, SE = 17.93, $t = -2.31$, $p = 0.03$) or Spillover 5 ($\widehat{\beta}$=69.23, SE = 29.59, $t = -2.34$, $p = 0.06$).

Table 4
Experiment 2 reading times (ms) by condition and by region (participant means ± standard error)

	Instrument	Spill1	Spill2	Spill3	Spill4	Spill5
Canonical Typical	467.53±18.16	410.48±12.21	395.91±12.69	404.43±12.72	412.24±14.11	488.58±21.44
Canonical Atypical	553.25±34.07	465.13±17.10	421.59±13.69	406.84±12.76	432.91±15.74	547.34±25.17
Cleft Typical	539.99±25.27	437.11±13.27	418.48±14.19	417.41±14.05	438.50±16.04	527.55±23.45
Cleft Atypical	555.93±27.94	511.86±16.01	423.53±12.77	415.86±12.07	428.30±13.08	512.39±18.35

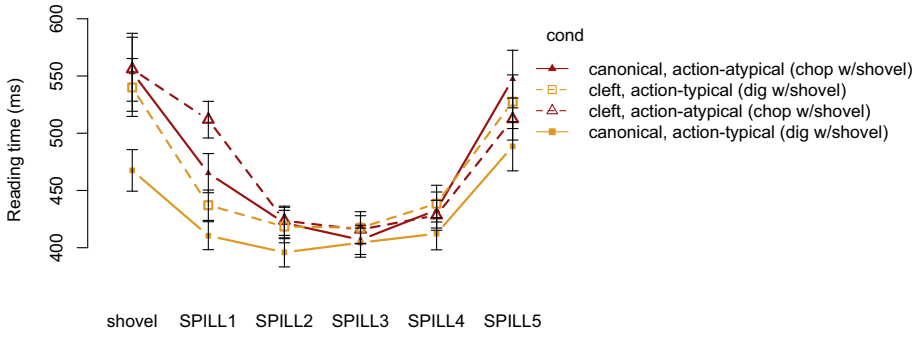


Fig. 3. Experiment 2 reading times (ms) by condition, from target instrument to end of sentence.

Table 5
Results of linear mixed-effect models for Experiment 2 reading time data. Boldface indicates significance after Bonferroni correction for the six regions' non-independent analyses (adjusted threshold =0.008)

Instrument	$\hat{\beta}$	SE	t	p	**Spillover 3**	$\hat{\beta}$	SE	t	p
(Intercept)	**531.30**	**23.05**	**23.06**	**<0.001**	**(Intercept)**	**413.12**	**11.14**	**37.07**	**<0.001**
Syntax	40.19	18.78	2.14	0.04	Syntax	10.42	11.04	0.94	0.37
Typicality	52.40	19.12	2.74	0.02	Typicality	4.21	9.52	0.44	0.67
Syntax × Typ	−74.06	32.57	−2.27	0.02	Syntax × Typ	−3.88	14.38	−0.27	0.79
Spillover 1	$\hat{\beta}$	SE	t	p	**Spillover 4**	$\hat{\beta}$	SE	t	p
(Intercept)	**455.78**	**12.80**	**35.62**	**<0.001**	**(Intercept)**	**430.09**	**13.46**	**31.96**	**<0.001**
Syntax	31.45	10.82	2.91	0.02	Syntax	14.07	12.31	1.14	0.28
Typicality	**63.17**	**14.61**	**4.32**	**<0.001**	Typicality	4.32	8.92	0.48	0.63
Syntax × Typ	10.68	21.01	0.51	0.62	Syntax × Typ	−30.53	15.75	−1.94	0.06
spillover 2	$\hat{\beta}$	SE	t	p	**spillover 5**	$\hat{\beta}$	SE	t	p
(Intercept)	**416.45**	**11.55**	**36.05**	**<0.001**	**(Intercept)**	**521.97**	**20.85**	**25.03**	**<0.001**
Syntax	14.10	14.72	0.96	0.36	Syntax	3.79	14.36	0.26	0.79
Typicality	12.91	7.79	1.66	0.10	Typicality	23.71	15.12	1.57	0.12
Syntax × Typ	−21.52	13.91	−1.55	0.12	**Syntax × Typ**	**−64.84**	**24.04**	**−2.70**	**0.008**

See Appendix C for the alternative single-model approach. In that analysis, the critical Syntax × Typicality interaction is significant, indicating that the interaction is present at the target word, the reference level for Region. The single-model approach thus shows the presence of the critical interaction at the earliest possible region, an effect that is not apparent after Bonferroni correction in the region-by-region analysis.

3.3. Discussion

Consistent with a comprehension bias in favor of informativity, participants showed a context-sensitive response to instrument typicality: The difficulty associated with reading about action-atypical information differs between the wh- cleft and no-cleft conditions. In the region-by-region analysis with Bonferroni corrected p-values, this critical Typicality × Syntax interaction is only significant sentence-finally (and is not present at all in the alternative analyses in footnote 5). In the alternative single-model analysis in Appendix C, the interaction is visible much earlier, at the target region, suggesting that incremental informativity-driven effects may be possible. Additional findings across analyses include main effects of Typicality, in keeping with known plausibility-driven effects, and Syntax, in keeping with similar findings for slowdowns with the clefted constituent in prior work on focus constructions.

In contrast to the results from Experiment 1, the reading times for Experiment 2 do not show the full reversal whereby the action-atypical instrument is read reliably faster than the action-typical instrument. However, the results do show that ease with the typical instrument is restricted to the canonical syntax condition, and that this restriction can be found as early as the target region as well as sentence finally. This reduction of the standard typicality effects is compatible with an account in which comprehension incorporates a bias in favor of informativity. It is also possible, however, that the complexity of the cleft structure makes it more difficult to track semantic relationships related to typicality. In that case, the reduction of the typicality effect could be attributed simply to the syntactic complexity of the cleft structure itself, not the information-structural constraints associated with it.

For the protagonist manipulation in Experiment 1, one could argue that the observed pattern of results shows that comprehenders can

simply shift their estimates of what is real-world predictable for a "normal" person to a different distribution over situations for a "surprising" person. In other words, it wasn't their expectations for newsworthiness that were determining their processing; rather they were adjusting their situation priors in a particular context. In contrast, the syntactic manipulation used in Experiment 2 can be said to more directly target comprehenders' linguistic expectations. The situations and individuals being described do not vary in terms of their real-world typicality; rather, it is the way the content is packaged syntactically that drives the effects. The final pair of experiments use neither surprising protagonists nor non-canonical structures; rather, we test if a speaker's intentional act of communicating by choosing to send a message to a comprehender is sufficient to influence comprehenders' expectations for newsworthiness.

4. Experiment 3a: Communication as a cue to unpredictability

Experiments 3a and 3b use text-message conversations as the context for the target items. The use of a natural dialogue setting is meant to evoke a speaker who is communicating intentionally. The speaker's intentional choice is emphasized in the dialogue via an initial exchange establishing that the speaker is making contact out of the blue ("hey, long time no see!") as a reminder of the cost/utility tradeoff of typing out a message and sharing news with a friend. The question is whether intentional communication is sufficient to guide expectations about message newsworthiness. If yes, we should see an informativity-driven effect even without any explicit cues to newsworthiness. If no, standard plausibility-driven effects should hold. We compare reading times for messages that describe typical and atypical situations. Again, we recruit a large number of participants each of whom see only one item per condition, with the aim of helping participants maintain their discourse expectations.

4.1. Methods

4.1.1. Participants

Four hundred forty six participants were recruited via Amazon Mechanical Turk and paid $0.80.

4.1.2. Materials

The experimental items were embedded in a text-message dialogue. The target region was always a number, either a situation-typical or situation-atypical value. One version of the dialogue is shown in (9). All target items appeared in the dialogue shown in (9), with four possible items in the first half of the dialogue (mall purchases) and four possible items in the second half (roommate promises). The items and target numeric values were counterbalanced so that a particular value was situation-typical in one dialogue and situation-atypical in another, as per Table 6.

(9)

JOE: hey, long time no see!
AMANDA: wow, a blast from the past!
AMANDA: how goes?
JOE: not much happening
JOE: but I thought of you today
JOE: I was at the mall
AMANDA: :) buying a present for me?!
Joe: :) no, but they had socks in your favorite shade of purple
Joe: I wanted to tell you what I saw – the price for the socks
was actually $2, that's what I saw! [situation-typical]
AMANDA: hey did you hear about the party on Friday? are you going?
JOE: what party?
AMANDA: it's at your old roommates' place
JOE: oh yeah?
JOE: we don't talk much anymore, not after the whole incident with the car and my cat
and so we kind of stopped talking after that
AMANDA: well they're throwing a party on Friday night and everyone's invited
JOE: don't go
AMANDA: I'm just gonna stop by, they promised me a drink
JOE: oh my god, the promises they make
Joe: once they promised they'd bake a dozen cookies
And then the actual number was 5, that was the number! [situation-atypical]

Table 6
Experiment 3a-b item/value combinations for target sentences

Item	Predictable Value	Unpredictable Value
Price of socks	$2	$150
Price leather coat	$150	$2
Price of headband	$10	$200
Price of Versace scarf	$200	$10
Bake a dozen cookies	12	5
Invite 5 people to a party	5	12
Age of young woman	25	5
Age of young child	5	25

As a pre-test of a preference for informativity, we elicited felicity judgments for the situation-typical and situation-atypical messages. Participants ($N = 31$ monolingual English speakers) were asked to rate on a scale of 1 to 7 how likely someone would be to send a text message about a situation, given that the situation had occurred: e.g., *If Joe saw socks for sale that cost $150, how likely is it that he would send you a message about it?* or *If Joe's friend promised to bake a dozen cookies and then baked 12, how likely is it that he would send you a message about it?* The results showed higher felicity ratings for messages with situation-atypical content (mean 4.77) than situation-typical content (mean 2.77; $\widehat{\beta} = -2.00$, SE = 0.32, $t = -6.24$, $p<0.001$). An item-by-item analysis showed this to be the case numerically for all items, with *t*-tests showing a significant difference for six pairs (socks, coat, headband, party, woman, child) but not for two (scarf, cookies). This pre-test helps address a potential concern that properties of the dialogues themselves (e.g., the untrustworthy roommates) might induce context-specific expectations. If that were the case, any observed reading time effects could reflect those properties rather than a more general expectation for informativity in speakers' text messages. However, the ratings suggest that the informative messages in our materials are generally favored over the uninformative ones, even with no dialogue context.

4.1.3. Procedure

Experiment 3 again uses a self-paced reading task. Participants were told they would be reading a text message conversation between two friends. Having indicated their consent, each participant saw two practice sentences, followed by the text-message dialogue presented one message at a time. The participant pressed the space bar to reveal each word non-cumulatively. Each dialogue contained only two critical items, one in each condition, with the order of the typical/atypical conditions counterbalanced across dialogues. The small number of critical items helps ensure that participants do not become accustomed to reading many discourse-infelicitous sentences. There were no comprehension questions that interrupted the dialogue but, to encourage participants to pay attention, they were told to expect a single comprehension question after the task was completed.

4.1.4. Analysis

The region-by-region analysis followed that of Experiments 1 and 2. We used linear mixed-effects modelling with a single fixed effect for Typicality (situation-typical vs. situation-atypical, coded as −0.5 and

0.5). For the by-items random effect structure, we include random intercepts and random slopes. For the by-participants random effect structure, we only include random intercepts because each participant only provides one datapoint for each level of the Typicality condition, failing to yield any variance for the model to capture with a by-participant random slope. Again, we use a Bonferroni-corrected threshold for determining significance. See Appendix D for the single-model approach. Both modelling approaches show the same pattern of results.

4.2. Results

The dataset for analysis here consists of reading times from 396 participants who were native English speakers who answered the task-final comprehension question correctly. In the resulting dataset, every item/condition combination was seen by at least 32 participants and an average of 60.3 participants. As a first cut, we removed reading times below 100 ms and above 5000 ms (66 datapoints in the critical and spillover regions). We then removed reading times more than three standard deviations away from the mean, per region and per condition (a further 63 datapoints in the regions of interest).

Table 7 and Fig. 4 show the raw reading times by condition for the critical region and the four spillover regions. Table 8 shows the region-by-region model output. After Bonferroni correction (adjusted threshold for significance =0.01), the results reveal the predicted effect of Typicality in the fourth spillover region, the final region of the sentence: Text messages about atypical values yielded faster sentence-final reading times than messages about typical values ($\hat{\beta} = -110.39$, $p < 0.001$). No other regions show significant effects before or after Bonferroni correction.

See Appendix D for the alternative single-model approach. The results show the critical interaction between Protagonist, Typicality, and Region (at Spill4), confirming that the interaction at the sentence-final region differs from that at the target region, the reference level for Region.

4.3. Discussion

Consistent with a comprehension bias in favor of informativity, participants showed faster reading times for sentences about real-world atypical situations compared to real-world typical situations (significant in the region-by-region analysis and the alternative analysis in Appendix D). The slower reading times for sentences about typical situations emerges at the final spillover region, as in Experiment 1 and 2. In this experiment, however, there is no contextual manipulation; it is presumably the global communicative intent of the speaker that raises comprehenders' expectation for something newsworthy.

That said, the target sentences in Experiment 3a did contain the words *actually* and *actual* before the critical region. In one sense, these words merely confirm that the speaker is telling the truth — something generally expected of speakers anyway. However, not all messages need confirmation of their truth, and it may be that speakers are more likely to include *actually* and *actual* when the truth of their message would otherwise be hard to believe. As such, these words could signal to comprehenders upcoming surprisal. Indeed, such words are known to emphasize novelty (Aijmer, 2013) and are associated with contrast and counter-expectation (Fetzer, 2009; Rohde et al., 2016; Simon-Vandenbergen & Aijmer, 2007). Experiment 3b replicates 3a without the words *actual(ly)*.

5. Experiment 3b: Communication as a cue to unpredictability, no 'actually'

The findings from Experiment 3a could be taken as evidence that expectations for informativity arise simply from encountering messages sent by an intentionally communicative speaker. Alternatively, the use of *actual(ly)* in the target sentences may have cued participants to expect situation-atypical content (analogous to Experiment 1's surprising protagonist or Experiment 2's wh- cleft construction). Here we remove the words *actually* and *actual*.

5.1. Methods

5.1.1. Participants

Four hundred eleven participants were recruited via Amazon Mechanical Turk and paid $0.80.

5.1.2. Materials

The text-message conversation was the same as in Experiment 3a, except that the words *actually* and *actual* were removed and a separate expression of emphasis was added after the fourth spillover region. The new message-final expression consisted of either *for real* or *honestly* (appearing as a single region), as in (10). These expressions were added to help make the conversation sound more natural, given the slightly repetitive phrasing of the target sentence. They technically do not contribute additional content beyond an affirmation by the speaker that the message is truthful, but such expressions (like *actually*) may also be used by speakers to acknowledge that a message is surprising enough to merit confirmation of its truth. Crucially, here they appear after the sentence-final spillover4 region where the Typicality effect was found in Experiment 3a.

(10)

a. [situation-typical]
I wanted to tell you what I saw – the price for the socks was $2, that's what I saw! for_real.

b. [situation-atypical]
I wanted to tell you what I saw – the price for the socks was $150, that's what I saw! for_real.

5.1.3. Procedure

The procedure followed that of Experiment 3a.

5.1.4. Analysis

The analysis followed that of Experiment 3a.

5.2. Results

The results represent data from 376 native-English-speaking participants who answered the task-final comprehension question correctly. Every item/condition combination was seen by at least 34 participants and an average of 57.1 participants. The first-cut outlier removal eliminated reading times below 100 ms and above 5000 ms (45 datapoints in the critical and spillover regions). We then removed reading times more than three standard deviations away from the mean, per region and per condition (a further 69 datapoints in the regions of interest).

Table 9 and Fig. 5 show the raw reading times by condition for the critical region and the five spillover regions. Table 10 shows the model results for each region. After Bonferroni correction (adjusted threshold

Table 7
Experiment 3a reading times (ms) by condition and by region (participant means ± standard error)

	Target	Spill1	Spill2	Spill3	Spill4
Situation-typical	509.53±15.63	485.19±11.50	415.49±7.77	411.84±7.19	755.61±27.80
Situation-atypical	518.70±16.53	517.95±13.31	409.07±7.44	412.41±8.15	649.84±21.17

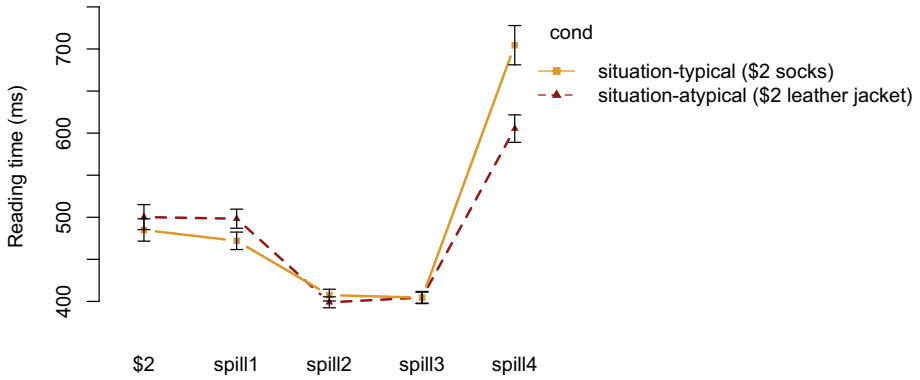


Fig. 4. Experiment 3a reading times (ms) by condition, from target region to end of sentence.

Table 8
Results of linear mixed-effect models of Experiment 3a reading time data. Boldface indicates significance after Bonferroni correction for the five regions' non-independent analyses (adjusted threshold =0.01)

Target	$\widehat{\beta}$	SE	t	p	**Spillover 3**	$\widehat{\beta}$	SE	t	p
(Intercept)	**527.75**	**21.32**	**24.75**	**<0.001**	**(Intercept)**	**413.53**	**11.80**	**35.04**	**<0.001**
Typicality	16.60	12.88	1.29	0.20	Typicality	1.01	6.56	0.15	0.88
Spillover 1	$\widehat{\beta}$	SE	t	p	**Spillover 4**	$\widehat{\beta}$	SE	t	p
(Intercept)	**511.20**	**16.37**	**31.23**	**<0.001**	**(Intercept)**	**651.21**	**26.14**	**24.92**	**<0.001**
Typicality	19.75	13.03	1.52	0.13	**Typicality**	**−110.39**	**23.36**	**−4.73**	**<0.001**
Spillover 2	$\widehat{\beta}$	SE	t	p					
(Intercept)	**413.16**	**10.80**	**38.27**	**<0.001**					
Typicality	−2.22	6.83	−0.33	0.75					

for significance =0.008), we see several main effects of Typicality. At the target and first spillover region, atypical situations are harder to read than typical situations (target: $\widehat{\beta}$=36.82, p<0.005; spillover1: $\widehat{\beta}$=67.63, p<0.001). This is similar to the numeric patterns of Experiments 1, 2, and 3a and in keeping with plausibility-driven effects. At the fourth spillover region, the pattern from Experiment 3a is replicated whereby atypical situations become easier to read than typical situations, but this effect does not survive Bonferroni correction.

See Appendix E for the alternative single-model approach. The results show the critical interaction between Protagonist, Typicality, and Region, confirming that the interactions at the sentence-final regions differ from that at the target region, the reference level for Region. The effect of Typicality at the target word is characterized by a significant slowdown for atypicality, whereas in Spill2, Spill3, Spill4, and Spill5 this slowdown for atypicality is reduced and significantly reversed at Spill4. As such, under the single-model approach, the predicted effect of Typicality at the fourth spillover region is significant.

5.3. Discussion

Experiment 3b is in keeping with the findings from Experiment 3a: At the sentence-final regions, situation-atypical content yielded faster reading times than situation-typical content (significant in the single-model approach reported in Appendix E but not in the region-by-region analysis with Bonferroni corrections). We take this as tentative confirmation that Experiment 3a's results extend to sentences without any explicit *actual(ly)* cue. The experiment also reveals early plausibility-driven effects whereby situation-atypical content yields slower reading times than situation-typical content at the target word and at the first spillover. In Experiment 3a, this pattern was present numerically at the same regions. It thus appears that comprehenders experience initial processing slowdowns when they encounter atypical content, but by the end of the utterance, it is the lack of any newsworthy or unpredictable content that yields slower reading times. We return to this point in the General Discussion.

A potential concern with the materials for Experiment 3b is that, even without the presence of *actual*, there may be other linguistic cues that signal that newsworthy content is coming. In (9), one target item starts with Joe saying *I wanted to tell you what I saw*. It may be that this sentence works to focus the listener on what is coming next by marking the importance of the upcoming content (*I want to tell you*) and by setting the Question Under Discussion (QUD) for upcoming discourse (*what I saw*). However, setting a QUD does not itself dictate an expectation for informativity because a QUD could be understood to merely signalthe topic of the upcoming discourse. The topic could be unfamiliar novel information (e.g., *I want to tell you what weird thing happened to me last night*) or familiar old information (e.g., *I want to remind you again what the doctor said*), which would suggest that QUDs are distinct from novelty/ informativity. Also, to the extent that Joe's phrasing sets up an expectation for an informative message (akin to *Hey, guess what?!*), such cues are rampant in natural discourse and their ubiquity underscores the point of this experiment, that one does not need a special context (like a surprising protagonist or a rare syntactic construction) to trigger comprehenders' bias in favor of informativity.

6. General discussion

The experiments reported here tested for informativity-driven expectations during processing. While comprehenders did show the well-established slowdowns for content about atypical situations, this slowdown was malleable. In contexts that highlighted the use of language as a medium for conveying informative and interesting messages, sentences about action-atypical instruments and other situation-atypical content yielded reading times that were as fast or faster than those for more typical content.

The results align with a model of informativity-driven processing in

Table 9
Experiment 3b reading times by condition and by region (participant means ± standard error)

	Target	Spill1	Spill2	Spill3	Spill4	Spill5
Typ	471.59±15.35	448.67±12.44	402.86±10.14	390.05±9.39	461.3087±13.37	749.26±27.59
Atyp	506.94±18.85	521.75±15.81	420.49±9.88	398.59±10.30	427.13±10.61	665.47±20.24

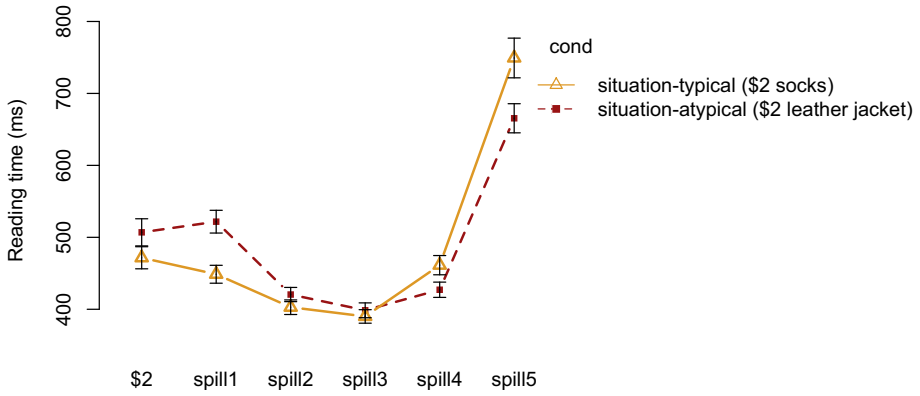


Fig. 5. Experiment 3b reading times (ms) by condition, from target number to end of sentence.

Table 10
Results of linear mixed-effect models of Experiment 3b reading time data. Boldface indicates significance after Bonferroni correction for the six regions' non-independent analyses (adjusted threshold =0.008)

Target	$\widehat{\beta}$	SE	*t*	*p*	**Spillover 3**	$\widehat{\beta}$	SE	*t*	*p*
(Intercept)	**494.32**	**20.03**	**24.68**	**<0.001**	**(Intercept)**	**395.86**	**8.07**	**49.09**	**<0.001**
Typicality	**36.82**	**12.95**	**2.84**	**<0.005**	Typicality	6.96	8.87	0.78	0.44
Spillover 1	$\widehat{\beta}$	SE	*t*	*p*	**Spillover 4**	$\widehat{\beta}$	SE	*t*	*p*
(Intercept)	**486.23**	**12.14**	**40.04**	**<0.001**	**(Intercept)**	**445.49**	**11.36**	**39.20**	**<0.001**
Typicality	**67.63**	**15.96**	**4.24**	**<0.001**	Typicality	−35.96	11.53	−3.12	0.03
Spillover 2	$\widehat{\beta}$	SE	*t*	*p*	**Spillover 5**	$\widehat{\beta}$	SE	*t*	*p*
(Intercept)	**414.03**	**8.10**	**51.09**	**<0.001**	**(Intercept)**	**710.27**	**20.77**	**34.20**	**<0.001**
Typicality	14.55	12.58	1.16	0.32	Typicality	−92.03	36.77	−2.50	0.13

which the probability comprehenders assign to a particular message combines two components — the estimated probability of the situation being described as well as the estimated probability that a speaker would choose to send a message about such a situation. A model that collapses situation typicality and utterance expectedness cannot account for the apparent difficulty our comprehenders show when they read about situations that are perfectly plausible. Our findings are compatible with accounts that describe comprehension as a process of reverse engineering a speaker's communicative intention. Not only do comprehenders need to update their mental model of the situation being described, but they also need to resolve why a speaker would have chosen to talk about that situation in the first place.

Our results stand in contrast to prior work that has shown facilitation for situation-typical content or Cloze-task predictable words; in our studies, the newsworthy content is neither situation-typical nor Cloze-task predictable but yet newsworthiness is shown to yield facilitation. This likely reflects our construction of contexts where informative messages are more expected, unlike in prior work. In addition, we took steps to prevent participants from losing their sense of what makes for felicitous discourse. Each participant saw only one item per condition so they did not experience a large number of infelicitous sentences, a problem in many studies which may influence reading behavior. In the last two studies, the sentences were embedded in a text-message dialogue to further maintain participants' normal discourse expectations and their bias for informativity if they had one.

Regarding the time course, our manipulations were intended to test whether informativity-driven effects are observable at all, and we saw effects primarily at the final region of the sentence. The potential exception to this was our use of a syntactic construction that helps signal the location of new information. With the wh- cleft construction in Experiment 2, effects can be seen at the target region itself in the single-model analysis approach that report in the Appendix. We take this as suggestive evidence that comprehension biases in favor of informativity can be deployed during moment-by-moment processing. Many sentences do not signal where the informative content will appear or even when the end of the sentence will arrive. In those cases, the lack of newsworthy content only becomes apparent at the very end of the sentence or when it is clear that a speaker has finished their turn. In other words, even comprehenders with high informativity expectations may give speakers the benefit of the doubt that some content is necessary to set the scene and that every word need not convey earth-shattering news. A speaker may start by describing a normal situation (*I was chopping carrots yesterday*), and if they stop there, a comprehender may be surprised by the message's lack of newsworthiness (or they will draw inferences about why *chopping carrots* or doing it *yesterday* could be interesting enough to merit commentary). But if the comprehender understands that the intended message may require some initial setup, they may accept some low-informativity scene setting as unproblematic before eventually being bothered if nothing newsworthy is ever delivered.

That said, processing newsworthy content is unlikely to be cost-free. Even if the position of informative content has been made clear and a comprehender knows that it's coming, they still have to grapple with the words themselves that convey this surprising content. In listening to a sentence (*I was chopping carrots*), a comprehender will presumably build a mental model of the events being described and will activate related concepts and words (knife, kitchen, salad, etc.). This automatic activation is useful in that it can facilitate processing if a subsequent sentence mentions one of those activated words (*I was standing in the kitchen holding the knife…*). But then when informative content finally arrives (e.g., *…when this bird flew in the window*), the very fact that it is informative means that it is not easily inferrable from context and the word or situation is thus less likely to have been pre-activated (e.g., *bird*). It may be the case that cues that strongly foreshadow upcoming informativity may allow comprehenders to suppress activation to some of the situation-typical concepts, although it still may be difficult for comprehenders to estimate which situation-atypical concepts are worth activating. In this sense, words that convey informative content may induce some processing difficulty when first encountered since they require extra work for lexical retrieval and instantiation in a mental model of the events being described, but these words may nonetheless yield a sentence that is preferred over a disappointingly uninformative one (e.g., *… when the kitchen clock ticked*). In our studies, this pattern is apparent in the way atypical content induces an initial slowdown at the target word or just after (numerically in Experiment 1 and 3a, significantly in Experiment 2 and 3b), followed by sentence-final facilitation (Experiments 1, 2, and 3).

Sentence-final reading responses have been discussed extensively in the literature (going back to Just & Carpenter, 1980, see recent review by Stowe, Kaan, Sabourin, & Taylor, 2018). Wrap-up effects are often characterized as a special sentence-final stage of processing that involves finalizing the syntax of the sentence and integrating the semantic proposition into the larger discourse context. In the ERP literature, brain responses at sentence-final regions have been attributed to processes related to syntactic wrap-up (because the ERP response to nonsense sentences differs from the response to word lists; Van Petten & Kutas, 1991) or to semantic integration (because it resembles the response typically present for semantic difficulties; Osterhout & Holcomb, 1992). However, there is also evidence that sentences with semantic disruption early on (for words with low Cloze probability or zero Cloze probability) still yield a sentence-final brain response that is similar to that of sentences with no disruption (Kutas & Hillyard, 1980; Van Petten & Kutas, 1991). These debates remain unresolved (see Warren, White, & Reichle, 2009), but they serve as a reminder of the relevance of sentence-final processing, which is particularly important for work on informativity if the sentence-final region is the primary position where the informativity of a message can be assessed.

A question we should perhaps be asking in psycholinguistics is whether our long-standing approach to studying predictability in language has over-emphasized comprehenders' situation knowledge (an impressively far-reaching and fine-grained knowledge system to be able to deploy in real-time) at the expense of comprehenders' expectations for what speakers actually try to communicate. The emphasis on speed may have led us down this path — real-time early-stage measures are often taken as a benchmark of importance for determining our psycholinguistic models, and that approach has indeed provided insight into the nature of the mental lexicon, the parser, and the neuroanatomy of language processing. But we shouldn't only be interested in what is easiest to understand in the first moments of processing. We also want models that account for the ensuing status of a sentence as it is interpreted at later stages (see the following for studies that grapple with such issues: Ferreira, Bailey, & Ferraro, 2002; Sanford & Sturt, 2002; Levy, Bicknell, Slattery, & Rayner, 2009) and within a pragmatic representation (Stewart, Haigh, & Ferguson, 2013). We see this as particularly relevant to computational applications like natural language generation and other domains in which representing globally 'good' sentences may depend on more than local word-by-word surprisal.

The approach advocated here raises new questions about what repercussions arise for listeners as they process overly predictable content. The results here show that they slow down, but are there specific inferences that arise when such material is encountered? One possibility is that comprehenders may adjust their estimate of what counts as newsworthy. For example, they may hear a speaker say *These socks cost $2* and conclude that the speaker must have a flatter likelihood distribution across messages of varying situation priors (i.e., this speaker just says out loud any observation, no matter how mundane). Or they may conclude that the audience is one for whom such information might really be informative (i.e., the addressee is a child). Another possibility is that comprehenders will decide to adjust their situation priors following the reasoning that cooperative speakers generally choose newsworthy messages — maybe the price of socks is going up and these particular socks are actually interestingly cheap! Recent work shows such effects: Narratives that explicitly mention inferable information change people's situation priors (Kravtchenko & Demberg, 2015). For example, in a narrative about someone shopping, the inclusion of an explicit statement that *He paid the cashier* causes participants to reduce their estimates of how often this individual normally pays. Such a finding points to the non-transparency of language use — under a transparent model in which real-world knowledge is directly linked to linguistic expression, announcing that a situation happened shouldn't reduce the estimate that the situation normally happens. Under an informativity-driven account, this result makes complete sense.

7. Conclusion

In contrast to work that highlights the role of situation typicality and word predictability in facilitating language comprehension, the studies we present here establish that comprehenders can also show relative ease with situation-atypical Cloze-task-unpredictable content. The effects are found primarily sentence-finally, which could suggest that comprehenders conduct a late-stage pragmatic assessment of the newsworthiness of a sentence or, relatedly, that it is only at the final word that a sentence's lack of newsworthiness becomes clear since there is no further content to rescue the sentence from infelicity. The findings are unexpected if one assumes that situations map transparently into speakers' utterances, an assumption that is implicit in many psycholinguistic models. We introduce an informativity-driven approach which posits that comprehenders distinguish content typicality from content mention, such that a 'good' utterance that is easy to process is one that balances content plausibility and novelty. Our focus on content selection goes beyond prior models of rational speaker-listener behavior by targeting propositional meaning instead of referential form. The findings depict a comprehender who expects a speaker to have something interesting to say.

Acknowledgements

Experiments 1 and 2 appeared in Richard Futrell's Masters thesis at Stanford University. We thank three anonymous reviewers for very helpful comments. This work was supported in part by a Leverhulme Trust Prize in Languages and Literatures to H. Rohde.

Appendix A. Experiment 1 & 2 filler items

1. A basketball court is the kind of place where even tall people can feel short. For example, a six-foot-three guy is dwarfed by the 7-ft players.
2. A flower shop is a delightful place filled with the best smells. For example, the one on my corner is filled with lilies at this time of year.
3. A hospital is the kind of building that needs a back-up generator. For instance, the operating room and the emergency room need to have power even during a storm.
4. Madonna is a celebrity who has been famous for decades. For example, kids growing up in the nineties knew her, as do kids today.
5. Mozart is a composer who everyone loves. For example, studies show that babies prefer Mozart over any other classical composer.
6. My aunt Sally is the kind of cook who never uses a recipe. For instance, yesterday she made a perfect souffle without even opening a cookbook.
7. My coworker Bob is the kind of athlete who practices obsessively. For instance, he wakes up at five thirty to go running every morning.
8. My uncle James is the kind of devoted father who has infinite patience. For instance, even on busy mornings he takes time to talk with each child about what they're going to do that day.
9. NYU is a trendy school where young actors and actresses often go to study. For example, Elizabeth Olsen studied there while filming movies on the side.
10. San Francisco is a hilly city with a lot great views. For example, from Potrero Hill you can see all of downtown.

Appendix B. Experiment 1 single-model analysis

Here we report an analysis approach which does not require Bonferroni correction because we avoid the multiple comparisons inherent to region-by-region analyses. Instead we construct a single model that includes three fixed effects: Protagonist (boring vs. surprising, coded as −0.5 and.5), Instrument Typicality, (action-typical vs. action-atypical, coded as −0.5 and.5), and Region (target word as reference level). We use maximal random effect structure as permitted by the data. An informativity-driven effect at the critical region would manifest as a 2-way Protagonist × Typicality interaction (showing the interaction holds at the reference level for Region, the target word). Later effects would manifest as 3-way Protagonist × Typicality × Region interactions. To clarify significant interactions in the omnibus model, we conduct follow-up analyses on the relevant subsets of the data.

Table A1 shows the model output for Experiment 1. The main effects of Region (Spill1, Spill3, and Spill4) indicate that the average reading time at those regions differs from that at the target region. We do not consider those effects further since the differences are not related to our experimental manipulations and likely arise from overall differences in reading speed for different length words or different sentence positions.

Table A1
Results of alternative single large model of Experiment 1 RT data; no Bonferroni corrections needed.

	$\widehat{\beta}$	SE	t	p
(Intercept)	**444.62**	**17.90**	**24.84**	**<0.001**
Protagonist	−30.24	19.85	−1.52	0.13
Typicality	2.95	19.66	0.15	0.88
Spill1	**35.03**	**13.86**	**2.53**	**0.01**
Spill2	−15.42	13.84	−1.11	0.27
Spill3	**−51.82**	**13.82**	**−3.75**	**<0.001**
Spill4	**228.00**	**13.91**	**16.39**	**<0.001**
Protagonist × Typicality	−11.02	39.29	−0.28	0.78
Protagonist × Spill1	14.68	27.73	0.53	0.60
Protagonist × Spill2	8.46	27.68	0.31	0.76
Protagonist × Spill3	21.73	27.63	0.79	0.43
Protagonist × Spill4	−14.10	27.77	−0.51	0.61
Typicality × Spill1	33.36	27.72	1.20	0.23
Typicality × Spill2	−5.46	27.68	−0.20	0.84
Typicality × Spill3	−3.00	27.63	−0.11	0.91
Typicality × Spill4	**−70.95**	**27.77**	**−2.56**	**0.01**
Protagonist × Typicality × Spill1	−34.28	55.46	−0.62	0.54
Protagonist × Typicality × Spill2	4.56	55.36	0.08	0.93
Protagonist × Typicality × Spill3	−12.77	55.26	−0.23	0.82
Protagonist × Typicality × Spill4	**−257.13**	**55.54**	**−4.63**	**<0.001**

Of primary interest are the interactions with our manipulated factors. There is a significant Typicality × Spill4 interaction, which indicates that Typicality has a different effect at Spill4 than it does at the target region. This is driven by the Protagonist × Typicality × Spill4 interaction. The follow-up analyses for these interactions correspond to the region-by-region analyses reported in the main text but with no Bonferroni corrections imposed. As shown in Table 3 in the main text, there is no main effect of Typicality at the target word (numerically, it shows longer reading times for atypical instruments) whereas there is a main effect of Typicality at Spill4 (whereby atypical instruments yield faster reading times). For the Protagonist × Typicality interaction; it is not significant at the target region but it is significant at Spill4, where it is driven in large part by the very fast reading times for the atypical instruments in the surprising protagonist condition.

In sum, the single-model analysis for Experiment 1 shows the predicted interaction at the sentence-final region.

Appendix C. Experiment 2 single-model analysis

The single-model analysis for Experiment 2 follows that for Experiment 1. Again, we construct a single model that includes three fixed effects: Syntax (canonical vs. cleft, coded as −0.5 and 0.5), Instrument Typicality, (action-typical vs. action-atypical, coded as −0.5 and 0.5), and Region (target word as reference level). As before, an informativity-driven effect at the critical region would manifest as a 2-way Syntax × Typicality interaction (showing that the interaction holds at the reference level for Region, the target word). Later-emerging effects would manifest as 3-way Syntax × Typicality × Region interactions.

Table A2 shows the model output for the overall model. As in Experiment 1, there are significant main effects of Region (Spill1, Spill2, Spill3, and Spill4), indicating that the average reading time in those regions differs from that at the target region. Again, what is of primary interest is the patterns with our manipulated factors. The results show a main effect of Syntax and a main effect of Typicality, whereby the cleft condition yields slower reading times than the canonical condition, and sentences with atypical instruments yield slower reading times than those with typical instruments. These main effects of Syntax and Typicality are driven by the predicted Syntax × Typicality interaction. Since the interaction does not interact with Region, it should be understood to hold at the reference level for Region, the target word. This significant Syntax × Typicality interaction thus provides evidence that informativity-driven effects can be localized to the target word.

To understand the direction of the Syntax × Typicality interaction at the target region, we conduct follow-up tests for a main effect of Typicality in the two Syntax conditions at that region: In the canonical condition, atypical instruments are read slower than typical instruments ($\widehat{\beta}$=86.31, SE = 27.37, $t = 3.15$, $p = 0.002$), whereas in the cleft condition, the difference is not significant ($\widehat{\beta}$=13.31, SE = 22.55, $t = 0.59$, $p = 0.56$). These follow-up tests were not licensed in the region-by-region analysis in the main text because the Syntax × Typicality interaction at the target region ($p = 0.02$) did not reach the Bonferroni-corrected threshold for significance ($p = 0.008$). The single-model approach reported here is not subject to Bonferroni correction and the significant interaction in the omnibus analysis is what licenses the follow-up tests.

Table A2
Results of alternative single large model of Experiment 2 RT data; no Bonferroni corrections needed.

	$\widehat{\beta}$	SE	t	p
(Intercept)	**529.43**	**14.63**	**36.20**	**<0.001**
Syntax	**38.72**	**14.06**	**2.75**	**0.008**
Typicality	**50.50**	**12.63**	**4.00**	**<0.001**
Spill1	**−71.82**	**8.74**	**−8.22**	**<0.001**
Spill2	**−112.51**	**8.71**	**−12.91**	**<0.001**
Spill3	**−115.68**	**8.72**	**−13.27**	**<0.001**
Spill4	**−99.52**	**8.71**	**−11.43**	**<0.001**
Spill5	−8.99	8.71	−1.03	0.30
Syntax × Typicality	**−71.86**	**24.70**	**−2.91**	**0.004**
Syntax × Spill1	−2.80	17.47	−0.16	0.87
Syntax × Spill2	−24.23	17.43	−1.39	0.16
Syntax × Spill3	−27.16	17.44	−1.56	0.12
Syntax × Spill4	−25.52	17.41	−1.47	0.14
Syntax × Spill5	**−34.59**	**17.41**	**−1.99**	**0.04**
Typicality × Spill1	14.97	17.46	0.86	0.39
Typicality × Spill2	**−35.41**	**17.43**	**−2.03**	**0.04**
Typicality × Spill3	**−45.95**	**17.44**	**−2.64**	**0.008**
Typicality × Spill4	**−46.98**	**17.41**	**−2.70**	**0.007**
Typicality × Spill5	−27.50	17.41	−1.58	0.11
Syntax × Typicality × Spill1	**88.87**	**34.93**	**2.54**	**0.01**
Syntax × Typicality × Spill2	48.71	34.85	1.40	0.16
Syntax × Typicality × Spill3	66.48	34.88	1.91	0.06
Syntax × Typicality × Spill4	40.72	34.83	1.17	0.24
Syntax × Typicality × Spill5	1.69	34.83	0.05	0.96

Table A2 also shows several other interactions, which together point to the difficulty with clefts and with atypical instruments, and offer evidence that helps localize the occurrence of the Syntax × Typicality interaction to the target word. The follow-up analyses in each case correspond to the region-by-region analyses reported in the main text (see Table 5), now without Bonferroni corrections.

First, the Syntax × Spill5 interaction shows that the slower reading times for cleft sentences differs between Spill5 and the target region. As reported in the region-by-region results in Table 5, the effect of Syntax (slower reading times for clefts than canonical structures) is marginal at the target region whereas it is non-significant at Spill5.

Second, Typicality interacts with Region, showing that the impact of Typicality at regions Spill2, Spill3, and Spill4 is different from its impact at the target word. As reported in the region-by-region results in Table 5, the effect of Typicality (slower reading times for atypical than typical instruments) is significant at the target word, whereas it is marginal or non-significant at Spill2, Spill3, and Spill4.

Lastly, the 3-way Syntax × Typicality × Spill1 interaction serves to localize the predicted 2-way Syntax × Typicality interaction to the target word by indicating that the 2-way interaction behaves differently at Spill1 compared to the reference level for Region. The region-by-region results in Table 5 show that the predicted Syntax × Typicality interaction is significant at the target region whereas it is not significant at Spill1. We also note the lack of a significant Syntax × Typicality × Spill5 interaction. The condition means at Spill5 indicate an interaction pattern that is similar to that at the target region. The lack of a significant difference between the behavior of the interaction at the target and Spill5 regions implies that a similar informativity-driven effect holds at the sentence-final region, which indeed matches the region-by-region analysis.

Appendix D. Experiment 3a single-model analysis

The single-model approach for Experiment 3a follows that for Experiments 1 and 2. We construct a single model that includes two fixed effects: Typicality (situation-typical vs. situation-atypical, coded as −0.5 and.5) and Region (target word as reference level). An informativity-driven effect at the critical region would manifest as a main effect of Typicality (showing the effect holds at the reference level for Region). Later effects would manifest as 2-way Typicality × Region interactions.

Table A3 shows the model output. As in Experiments 1 and 2, there are significant main effects of Region (Spill2, Spill3, and Spill4), indicating that the average reading time in those regions differs from that at the target region. In addition, the Typicality × Spill4 interaction is significant, showing that the effect of Typicality differs between Spill4 and the target region. For the follow-up to this interaction, we consult the region-by-region results in Table 8 in the main text. The effect of Typicality (whereby atypical values yield numerically slower reading times than typical values) is non-significant at the target word but is reversed and significant at the Spill4 (where atypical values yield faster reading times than typical ones).

Table A3
Results of alternative single large model of Experiment 3a RT data; no Bonferroni corrections needed.

	$\widehat{\beta}$	SE	t	p
(Intercept)	**523.24**	**16.80**	**31.15**	**<0.001**
Typicality	14.65	27.64	0.53	0.62
Spill1	−26.72	14.19	−1.88	0.11
Spill2	**−105.39**	**12.27**	**−8.59**	**<0.001**
Spill3	**−119.36**	**11.97**	**−9.97**	**<0.001**
Spill4	**192.47**	**23.89**	**8.06**	**<0.001**
Typicality × Spill1	−4.15	24.43	−0.17	0.87
Typicality × Spill2	−14.73	27.38	−0.54	0.60
Typicality × Spill3	−15.57	34.20	−0.46	0.66
Typicality × Spill4	**−129.48**	**44.29**	**−2.92**	**0.02**

Appendix E. Experiment 3b single-model analysis

The single-model approach for Experiment 3b matches that for Experiment 3a. Table A4 shows the model output. As before, there are significant main effects of Region (Spill2, Spill3, Spill4, and Spill5), indicating that the average reading time in those regions differs from that at the reference level, the target region. There are several Typicality × Region interactions which likewise indicate that the effect of Typicality is different in these regions compared to the target region. For the follow-ups to these interactions, we consult the region-by-region results in Table 10 in the main text. Those region-by-region analyses were discussed in the main text with Bonferroni corrections applied, whereas here no Bonferroni correction is needed. The effect of Typicality at the target word is characterized by significantly slower reading times for atypical situations compared to typical situations, whereas in Spill2, Spill3, Spill4, and Spill5 this pattern is reduced and then significantly reversed at Spill4. This finding is in keeping with the informativity-driven prediction that, at some point during processing, atypical situations can yield faster reading times than typical ones.

Table A4
Results of alternative single large model of Experiment 3b RT data; no Bonferroni corrections needed.

	$\widehat{\beta}$	SE	t	p
(Intercept)	**494.10**	**17.68**	**27.94**	**<0.001**
Typicality	33.40	24.44	1.37	0.24
Spill1	−3.20	19.43	−0.17	0.87
Spill2	**−78.01**	**18.71**	**−4.17**	**<0.005**
Spill3	**−98.94**	**17.53**	**−5.64**	**<0.001**
Spill4	**−42.77**	**15.41**	**−2.78**	**<0.05**
Spill5	**219.06**	**16.14**	**13.57**	**<0.001**
Typicality:Spill1	−4.96	24.07	−0.21	0.84
Typicality:Spill2	**−53.65**	**24.07**	**−2.23**	**<0.05**
Typicality:Spill3	**−70.39**	**24.03**	**−2.93**	**<0.005**
Typicality:Spill4	**−65.19**	**21.43**	**−3.04**	**<0.005**
Typicality:Spill5	**−120.10**	**21.54**	**−5.58**	**<0.001**

References

Aijmer, K. (2013). *Understanding pragmatic markers*. Edinburgh: Edinburgh University Press.

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review, 106*, 748–765.

Arnold, J. E., Tanenhaus, M. K., Altmann, R., & Fagnano, M. (2004). The old and thee, uh, new. *Psychological Science, 15*, 658–668.

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

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

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

Bennett, E., & Goodman, N. (2018). Extremely costly intensifiers are stronger than quite costly ones. *Cognition, 178*, 147–161.

Benz, A., Jäger, G., & Rooij, R. V. (2005). *Game theory and pragmatics*. Hampshire, UK: Palgrave Macmillan.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language, 63*, 489–505.

Birner, B. (2004). Discourse functions at the periphery: Non-canonical word order in english. In B. Shaer, W. Frey, & C. Maienborn (Eds.), *Proceedings of the dislocated elements workshop, ZAS papers in linguistics* (pp. 42–62). Berlin: ZAS.

Birner, B. J., & Ward, G. (2009). Information structure and syntactic structure. *Language and Linguistic Compass, 3*, 1167–1187.

Borovsky, A. (2017). The amount and structure of prior event experience affects anticipatory sentence interpretation. *Language, Cognition and Neuroscience, 32*, 190–204.

Boudewyn, M. A., Long, D. L., & Swaab, T. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension. *Cognitive, Affective, and Behavioral Neuroscience, 15*, 607–624.

Brants, T., & Franz, A. (2006). *Web 1T 5-gram version 1*.

Brown, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology, 19*, 441–472.

Connell, L., & Keane, M. T. (2004). What plausibly affects plausibility? Concept coherence and distributional word coherence as factors influencing plausibility judgments. *Memory and Cognition, 32*, 185–197.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition, 105*, 658–668.

Cowles, H. W., & Garnham, A. (2005). Antecedent focus and conceptual distance effects in category noun-phrase anaphora. *Language and Cognitive Processes, 20*, 725–750.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expression. *Cognitive Science, 18*, 233–263.

Davies, C., & Arnold, J. E. (2019). Reference and informativeness: How context shapes referential choice. In C. Cummins, & N. Katsos (Eds.), *Handbook of experimental semantics and pragmatics* (pp. 474–493). Oxford: Oxford University Press.

Degen, J., & Franke, M. (2012). Optimal reasoning about referential expressions. In S. Brown-Schmidt, J. Ginzburg, & S. Larsson (Eds.), *Proceedings of the 16th workshop on the semantics and pragmatics of dialogue* (pp. 2–11).

Degen, J., Hawkins, R., Graf, C., Kreiss, E., & Goodman, N. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review, 127*(4), 591–621.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the 37th annual conference of the Cognitive Science Society*.

Drummond, A. (2013). Ibex farm. http://spellout.net/ibexfarm.

Elman, J. L., & McRae, K. (2017). A model of event knowledge. *Psychology Publications, 122*, 1–6.

Ferguson, H. J., & Breheny, R. (2011). Eye movements reveal the time-course of anticipating behaviour based on complex, conflicting desires. *Cognition, 119*, 179–196.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*, 11–15.

Fetzer, A. (2009). Challenges in contrast: A function-to-form approach. In K. Aijmer (Ed.), *Contrastive pragmatics* (pp. 73–96). Amsterdam & Philadelphia: John Benjamins.

Filik, R., & Leuthold, H. (2008). Processing local pragmatic anomalies in fictional contexts: Evidence from the N400. *Psychophysiology, 45*, 554–558.

Fincher-Kiefer, R. (1996). Encoding differences between bridging and predictive inferences. *Discourse Processes, 22*, 225–246.

Fischler, I., Bloom, P., Childers, D., Roucos, S., & Perry, N. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology*, 20.

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *30th annual meeting of the Cognitive Science Society* (pp. 939–944).

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

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft, 35*, 3–44.

Gahl, S. (2008). "Time" and "Thyme" are not homophones: Word durations in spontaneous speech. *Language, 84*, 474–496.

Garnham, A., Traxler, M., Oakhill, J., & Gernsabacher, M. A. (1996). The locus of implicit causality effects in comprehension. *Journal of Memory and Language, 35*, 517–543.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Science, 110*, 8051–8056.

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

Gries, S. T., & Divjak, D. S. (2012). *Frequency effects in language: Learning and processing*. Berlin & New York: Morton de Gruyter.

Grigoroglou, M., & Papafragou, A. (2016). Are children flexible speakers? Effects of typicality and listener needs in children's event descriptions. In *Proceedings of the 38th annual meeting of the Cognitive Science Society* (pp. 782–787).

Grimes-Maguire, R., & Keane, M. T. (2005). Expecting a surprise? The effect of expectations in perceived surprise in stories. In *Proceedings of the 27th annual conference of the Cognitive Science Society. Erlbaum, Hillsdale, NJ* (pp. 833–838).

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

Grüter, T., Takeda, A., Rohde, H., & Schafer, A. (2018). Intersentential coreference expectations reflect mental models of events. *Cognition, 177*, 172–176.

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

Hald, L. A., Steenbeek-Planting, E. G., & Hagoort, P. (2007). The interaction of discourse context and world knowledge in online sentence comprehension: Evidence from the N400. *Brain Research, 1146*, 210–218.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language, 49*, 43–61.

Heller, D., Arnold, J. E., Klein, N., & Tanenhaus, M. K. (2015). Inferring difficulty: Flexibility in the real-time processing of disfluency. *Language and Speech, 58*, 190–203.

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

Jiang, X., Li, Y., & Zhou, X. (2013). Even a rich man can afford that expensive house: ERP responses to construction-based pragmatic constraints during sentence comprehension. *Neuropsychologia, 51*, 1857–1866.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*, 329–354.

Köhne-Fuetterer, J., Drenhaus, H., & Delogu, F. (2020). The online processing of causal and concessive discourse connectives. *Linguistics* (in press, Special Issue on Discourse Expectations).

Krahmer, E., & Deemter, K. V. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics, 38*, 173–218.

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

Kuperberg, G. (2016). Separate streams or probabilistic inference? What the n400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience, 31*, 602–616.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*, 32–59.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*, 203–205.

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

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences, 106*, 21086–21090.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of the 20th conference on neural information processing systems (NIPS), pp* (pp. 849–856).

Lewis, D. (1969). *Convention*. Harvard, MA: Harvard University Press.

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin and Review, 9*, 550–557.

Lowder, M. W., & Gordon, P. C. (2015). Focus takes time: Structural effects on reading. *Psychonomic Bulletin and Review, 22*, 1733–1738.

Marks, L. E., & Miller, G. A. (1964). The role of semantic and syntactic constraints in the memorization of english sentences. *Journal of Verbal Learning and Verbal Behavior, 3*, 1–5.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 37*, 913–934.

McKoon, G., & Ratcliff, R. (1986). Inferences about predictable events. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 12*, 82–91.

McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistic Compass, 3*, 1417–1429.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38*, 283–312.

Mitchell, M., Reiter, E., & Van Deemter, K. (2013). Typicality and object reference. In *Proceedings of the 35th annual meeting of the Cognitive Science Society* (pp. 3062–3067).

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 20*, 92–103.

Nieuwland, M. S. (2015). The truth before and after: Brain potentials reveal automatic activation of event knowledge during sentence comprehension. *Journal of Cognitive Neuroscience, 27*, 2215–2228.

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle. *Psychological Science, 19*, 1213–1218.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience, 18*, 1098–1111.

Nordmeyer, A. E., & Frank, M. C. (2015). The pragmatics of negation across contexts. In *Proceedings of the 37th annual conference of the Cognitive Science Society* (pp. 1739–1744).

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language, 31*, 785–806.

Pechman, T. (1989). Incremental speech production and referential overspecification. *Linguistics, 27*, 89–110.

Prince, E. (1978). A comparison of wh-clefts and it-clefts in discourse. *Language, 54*, 883–906.

R Core Team. (2017). R: A language and environment for statistical computing. In *R foundation for statistical computing. Vienna, Austria*. https://www.R-project.org/.

Rapp, D. N., & Gerrig, R. J. (2002). Readers reality-driven and plot-driven analyses in narrative comprehension. *Memory and Cognition, 30*, 779–788.

Rodríguez-Gómez, P., Sánchez-Carmona, A., Smith, C., Pozo, M. A., Hinojosa, J. A., & Moreno, E. M. (2016). On the violation of causal, emotional, and locative inferences: An event-related potentials study. *Neuropsychologia, 87*, 25–34.

Rohde, H., Dickinson, A., Schneider, N., Clark, C. N. L., Louis, A., & Webber, B. (2016). Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of 10th linguistic annotation workshop (LAW)* (pp. 49–58).

Rohde, H., Levy, R., & Kehler, A. (2011). Anticipating explanations in relative clause processing. *Cognition, 118*, 339–358.

Rohde, H., Seyfarth, S., Clark, B., Jaeger, G., & Kaufmann, S. (2012). Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In S. Brown-Schmidt, J. Ginzburg, & S. Larsson (Eds.), *Proceedings of the 16th workshop on the semantics and pragmatics of dialogue* (pp. 107–116).

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics, 1*, 75–116.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology, 7*, 1–15.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences, 6*, 382–386.

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

Simon-Vandenbergen, A. M., & Aijmer, K. (2007). *The semantic field of modal certainty. A corpus-based study of English adverbs*. Berlin/New York: Mouton de Gruyter.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. Oxford: Blackwell.

Stanovich, K. E., & West, R. F. (1979). Mechanisms of sentence context effects in reading: Automatic activation and conscious attention. *Memory and Cognition, 7*, 77–85.

Stewart, A. J., Haigh, M., & Ferguson, H. (2013). Sensitivity to speaker control in the online comprehension of conditional tips and promises: An eye tracking study. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 39*, 1022–1036.

Stewart, A. J., Pickering, M. J., & Sanford, A. J. (2000). The time course of the influence of implicit causality information: Focusing versus integration accounts. *Journal of Memory and Language, 42*, 423–443.

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Language Learning and Development, 11*, 176–190.

Stowe, L. A., Kaan, E., Sabourin, L., & Taylor, R. C. (2018). The sentence wrap-up dogma. *Cognition, 176*, 232–247.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*, 415–433.

Tian, Y., Breheny, R., & Ferguson, H. J. (2010). Why we simulate negated information: A dynamic pragmatic account. *The Quarterly Journal of Experimental Psychology, 63*, 2305–2312.

Troyer, M., & Kutas, M. (2018). Harry Potter and the chamber of what?: The impact of what individuals know on word processing during reading. *Language, Cognition, and Neuroscience*, 1–17.

Van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience, 20*, 580–591.

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory and Cognition, 19*, 95–112.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes, 56*, 229–255.

Walker, J. H. (1975). Real-world variability, reasonableness judgments, and memory representations for concepts. *Journal of Verbal Learning and Verbal Behavior, 14*, 241–252.

Warren, T., McConnell, K., & Rayner, K. (2008). Effects of context on eye movements when reading about possible and impossible events. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34*, 1001–1010.

Warren, T., Milburn, E. A., Patson, N. D., & Dickey, M. W. (2015). Comprehending the impossible: What role do selectional restriction violations play? *Language, Cognition, and Neuroscience, 30*, 932–939.

Warren, T., White, S. J., & Reichle, E. D. (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and E-Z reader. *Cognition, 111*, 132–137.

Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology, 6*, 1–12.

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience, 30*, 648–672.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the 38th annual meeting of the Cognitive Science Society* (pp. 2771–2776).

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162–185.