Why are you telling me this? The availability and timing of relevance inferences

Abstract

Part of successful communication involves recognising the purpose of, or the intentions underlying, what speakers choose to say. Often, such pragmatic inferences are studied with an emphasis on informativity. The present work however moves beyond the types of inferences typically studied in prior work and instead investigates inferences from more naturalistic utterances, specifically those whose triviality may invite addressees to reason about why a speaker would have made such a discourse contribution. We present four studies (total N=777) using offline and online methods to investigate how and when listeners derive *relevance inferences* from trivial utterances. We manipulate speaker knowledge, speaker style, and linguistic properties of the utterances to show that, even in the absence of explicit emphasis cues, trivial utterances such as "the library walls are blue" are likely to be understood as conveying more than what is stated explicitly (e.g. *that the walls used to be a different colour*), and that these inferences are more likely to arise when produced by a speaker who is knowledgeable about the situation and who does not typically talk a lot. Our results suggest that comprehenders have pervasive expectations of cooperativity which, when seemingly violated by a speaker's trivial utterance, prompt reasoning about a speaker's motivation for speaking to determine how the communicated content is relevant. We then turn to the processing costs of computing triviality-driven inferences and find evidence that there may be a cost to deriving relevance inferences. These findings extend previous work on inferencing, which typically targets specific classes of words that give rise to inferences and demonstrates that broader, systematic inferencing that can arise when addressees reason about speaker goals even in the absence of cues to pragmatic enrichment.

Part of communicating successfully involves recognising the intentions underlying what speakers choose to say (Grice, 1975; Levinson, 2000). While this recognition seems straightforward in that listeners map the incoming signal onto words and meaning, language is not always produced in such a direct or transparent manner that a speaker's utterance and intended meaning go hand in hand. Oftentimes, language is used indirectly and utterances do not transparently map onto the intended meaning. In these cases, successful communication requires listeners to go beyond the transparent mapping and infer additional meaning in order to determine how the utterance is relevant to the discourse. One way such inference is achieved is by reasoning about a speaker's communicative goals. When successful, indirect communication presents an efficient way for speakers to communicate an array of context-sensitive meanings in finite time. However, the availability of such inferences presents listeners with a challenge; when should they engage in the search for additional meaning and draw an inference?

In the present work we examine the availability of relevance inferences arising from *trivial utterances,* utterances that are neither blatantly underinformative (failing to provide a sufficient Quantity of information per Grice, 1975) nor explicitly redundant ("overinformative" per Rubio-Fernandez, 2019). We use trivial utterances as a way of testing the interpretation of a type of utterance that appears in natural communication. Pragmatic inferencing, typically, is studied couched in terms of *informativity* based on world knowledge, entailment relations, and referential communication. *Trivial utterances* offer the opportunity to examine a broader variety of utterances that prompt inferencing beyond the informational value of a description (how much information is contained in an utterance). Informational value emphasizes the amount of information and can be understood in a variety of ways: information as reduction of uncertainty about the world (Levinson, 2000; Shannon, 1948) or as an utterance's discriminatory value e.g. modification (*the banana* vs *the yellow banana* vs *the orange banana*; see Sedivy, 2003; Rubio-Fernandez, 2019), or the specificity of a label (*the animal, the bird, the crow;* Brennan & Clark, 1996; Engelhardt et al., 2006; see also Davies & Arnold, 2019). In contrast, consider a speaker who utters the following as commentary on a situation, without any preceding question inquiring about this information:

(1) The library walls are blue.

Upon hearing such an utterance, an addressee is faced with a choice to (a) interpret the utterance as transparently informing them about the state of the walls (i.e. what was said: "The library walls are blue") or (b) reason about the speaker's goals in producing such an utterance to determine how it is relevant; perhaps the speaker noticed a change (i.e. what was meant: *the library used to be a different colour*) or the speaker was surprised at the choice (i.e. what was meant: *the library walls are an unusual colour*). We refer to the additional meaning arising in (b) as a *relevance inference*, in the sense that a listener reasons about *why* a speaker has made the discourse contribution that they have.

Relevance inferences are prompted by an addressee reasoning about a speaker's choice to produce an utterance rather than say nothing. By creating an overt signal, the utterance carries an assumption or expectation of relevance. We assume that utterances produced are relevant to the current situation and that determining how they are relevant involves reasoning about the speaker's intention (Sperber & Wilson, 1995/2004; Wearing, 2015).

Consider the example from Grice (1975, p.51):

(2) A is standing by an obviously immobile car and is approached by B;
A: I am out of petrol.
B: There is a garage round the corner.

In this exchange, B's contribution is only a relevant contribution if they believe that the garage is open and has petrol. These additional elements of meaning are not explicitly stated and must be inferred through reasoning about *why* B chose to utter such an utterance. Without the assumption of relevance and without the inference of the additional meaning that the garage is open and has petrol, B's utterance is not a cooperative response in this dialogue.

To the extent that comprehenders draw relevance inferences, it is an open question as to the availability of these inferences during language comprehension. Here we present four studies using offline and online measures to investigate the computation of these inferences. We first examine linguistic and extralinguistic cues that may signal to comprehenders that an inference can or ought to be computed. We then probe comprehension effort both during the processing of the sentence itself (sentence reading times) and at a later time point when the participant is asked whether they endorse the inference meaning (response reaction times). Our results show that, as with other types of pragmatic inferencing, whether relevance inferences arise depends on extralinguistic cues such as speaker-specific characteristics (e.g. their knowledge and their speech style). We show these inferences arise most easily when a speaker is understood to be knowledgeable about the situation they are talking about and when the speaker is understood to be someone whose communicative choices reflect a preference to filter what they say.

In the processing data, we see a more nuanced picture: Inferring additional meaning appears to require a small increase in processing effort when processing the sentence itself, however, this finding requires further investigation. We see a clearer cost after the sentence has been processed when participants are prompted to endorse or refute the additional meaning. These findings parallel other work on pragmatic inferences that shows that the inference of additional meaning is readily available to comprehenders but not necessarily cost-free. The experiments presented here provide an avenue for broadening our understanding of pragmatic inferencing beyond traditionally studied inferences.

Using language

Language use is often characterised as cooperative joint action between interlocutors (Garrod & Pickering, 2009; Grice, 1975; Searle, 1992). Successful communication involves a speaker selecting the appropriate signals to convey their intended meaning and an addressee recovering that intended meaning (Geurts, 2010; Grice, 1975). When a speaker utters "Can you open the window?", the addressee is faced with a choice – interpret the utterance transparently and respond to the inquiry about their *ability* to open the window or recognise that the speaker is intending to be understood as making a request and therefore open the window. Conventionally, questions in English such as "Can you X?" are understood as requests, rather than as inquiries into ability. However, under certain circumstances the question may be interpreted as an inquiry about ability; for example, if the addressee has recently broken their arm and the speaker is genuinely asking if they are able to perform the action or not. Crucially, the success of this communicative interaction hinges on correctly retrieving the speaker's intended goal. It is often assumed that the goal of a speaker is to convey information or prompt their addressee to action. However, there are instances of phatic communication where language is used to establish and maintain social relationships rather than convey information

or ideas (Crystal, 1991; Laver, 1974; Malinowski, 1923; Žegarac & Clark, 1999). Consider the example (3):

(3) It's hot in here.

A speaker who utters (3) could be communicating phatically, without a goal or agenda beyond avoiding an awkward silence and offering an observation as small talk. However, it is also possible that this utterance is intended non-phatically and the speaker might be specifically conveying the information that it is hot in the room. In that case, (3) would be a redundant utterance if the speaker and addressee are co-located, since presumably they would both be experiencing the same temperature in the room. Thus, an addressee may infer a further goal on the speaker's part to get a window opened or a fan switched on. This ambiguity presents somewhat of a challenge to the addressee to identify the goal of the speaker in producing such an utterance, and the eventual interpretation may involve computing an inference.

The availability of indirect meaning raises the question of when comprehenders decide to go beyond transparent meaning and make an inference. One salient cue is an assessment that the utterance fails to meet their expectations for what constitutes an appropriate conversational contribution. There are a number of theories about what these specific expectations are (e.g. Geurts, 2010; Grice, 1975; Levinson, 2000; Bohn, Tessler, & Goodman, 2019). The central expectation is that our conversational partner is communicating rationally—namely, that they care about the goals of the exchange and formulate their contributions in a manner that allows them to communicate their message (Grice, 1975). Specifically, a rational speaker's utterance should be relevant, informative, produced simply, and convey all necessary information for the conversational exchange (e.g. Bohn, Tessler, & Goodman, 2019; Brown & Dell, 1987; Grice, 1975; Horn, 2004; Levinson, 2000; Sperber & Wilson, 1995/2004).

Pragmatic inferencing is often studied in terms of rational reasoning regarding the quantity of information stated or implied by an utterance; comprehenders expect an appropriate *amount* of information to be conveyed (see Quantity maxim Grice, 1975; Q- & R-principles Horn, 2004; and Q heuristic Levinson, 2000). If the message is deemed to have an insufficient quantity of information, a listener a listener may search for additional enrichments that would ensure that the utterance meets the expectations of cooperative discourse. Consider scalar implicatures, a type of *quantity,* or *informativity-based* inference, which can arise when a speaker has been seemingly underinformative by choosing not to use a stronger, more informative, expression. For example, a listener who hears *some students passed* or *the tea is warm* may draw an inference. The expressions *some* and *warm* are both members of lexical scales ordered on informativity (<some, all> and <warm, hot>) and their use is compatible with stronger meanings (*some and possibly all*, *warm and possibly hot*). When a speaker uses a less informative expression from the scale, listeners try to reconcile their expectations regarding speaker informativity with the seeming underinformativeness of the expression by inferring that the stronger (more informative) expression does not hold, thereby deriving a scalar implicature (Gazdar, 1979; Geurts, 2010; Grice, 1975; Horn, 1972, 1989; Levinson, 2000).

A common feature of scalar inferences is that there are specific lexical triggers whose underinformativeness prompts the derivation of those inferences.[1] In the case of scalar implicatures, it is often a quantifier such as "some" (see van Tiel et al., 2016). Models of pragmatic inferencing suggest that, providing the speaker is cooperative, the same scalar

[1] There are also contextually based scalar implicatures in the case of ad hoc scales (Hirschberg, 1985)

inferences would likely arise regardless of *who* specifically produces the utterance. In other words, the specific goals of a specific speaker are not what drives the emergence of scalar inferences. Rather, these emerge due to the inherent informativity of the lexical items used. Scalar implicature is a reliable but not specifically creative type of inference, in that it is often context independent. However, our linguistic experience is rich and many of the inferences that arise in conversations are not limited to the use of scalar terms that invoke reasoning about speaker underinformativeness.

Utterances need not be only *lacking* in information for comprehenders to compute an inference. In a series of studies, Kravtchenko and Demberg (2022) investigated pragmatic inferencing from overly informative, or redundant, propositions where the quantity of information provided was in excess of what would be expected. Through exploiting script knowledge of common events (e.g. going to a restaurant), they found that when a narrative explicitly mentioned an action that is easily inferred from the context (i.e. eating in a restaurant), comprehenders were found to interpret this action as atypical for the protagonist in the narrative (i.e. that the character does not typically eat when going to a restaurant). Upon encountering informationally redundant utterances, comprehenders worked to infer the speaker's communicative goals, leading to inferences that the action was mentioned *because* it was atypical for this protagonist and hence newsworthy (a so-called atypicality inference). That is, the speaker's choice to produce the utterance at all, given the obvious nature of the content, is often interpreted by listeners as a reflection of that feature's atypicality and a signal for an addressee to compute an inference (e.g. Bergey & Yurovsky, 2022; Horowitz & Frank, 2016; Kreiss & Degen, 2020; Ryzhova, Mayn & Demberg, 2023). Thus, the decision to say anything at all may be a strong enough cue for a comprehender to expect a suitably interesting or informative contribution; when the utterance fails to meet these criteria, additional meaning may be computed through reasoning about the speaker's goals.

Along with conveying a certain amount of information, speakers are also expected to convey this information in an appropriate way. Consider Levinson's (2000) I-principle, *what is not said is the obvious*, and M-principle, *what is said in an abnormal way is not normal*, which apply to the use of (4) and (5).

(4) Bill caused the car to stop.
(5) Bill stopped the car.

Under the M-principle, (4) implies that the car was stopped in a non-stereotypical way. Choosing to express a stereotypical event in this manner is not the conventional formulation. The marked nature of the utterance can lead addressees to try to reason why the utterance was produced in an unconventional manner when a more conventional alternative exists as in (5). Perhaps the speaker is trying to convey that Bill did not stop the car in the conventional manner e.g. rather than using the brakes, he used a tree.

Under the I-principle, we do not expect speakers to state the obvious or tell us information that is easily inferable. The choice to utter (5) is also marked since the event is stereotypical and the utterance can imply that, although the car was stopped stereotypically, there was a particular reason for doing so (as it is not usually noteworthy to mention that a car was stopped unless the object of interest is the cause, e.g. Bill stopped the car because the police flagged him down).

A core assumption within the literature on inferences that arise from apparent violations of manner or informativity is that the utterances are somehow *relevant* to the discourse. Without

this assumption, it is unclear why a listener would ever go to the trouble of computing an inference. For scalar implicatures, reconciling the violation of informativity involves reasoning about why a speaker chose not to utter the stronger alternative and therefore how the utterance produced is relevant to the current discourse. For redundant utterances or those with unconventional formulations, reconciling the seeming violation of conversational norms involves reasoning about why the speaker produced so much extra information or why they used extra words to communicate it (yielding atypicality inferences). In each of these instances, a comprehender is reasoning about how the utterance is relevant as stated; why did the speaker choose to say what they said?

Why did they say that?

Successfully communicating requires: (a) the production of a signal, (b) the intention of a speaker to inform an addressee of something, and (c) an addressee recognising the speaker's intention to inform . Thus, by choosing to produce an utterance (c) is realised and producing any utterance comes with a presumption of its own relevance (Sperber & Wilson, 1995/2004; Wearing, 2015). How an utterance is relevant is not always immediately clear.

Consider the following utterances:

(6) The library has books.

(7) The library opens at 9:30am.

(8) The library walls are blue.

Each of these utterances conveys some information about the library, but they vary in their informativity or usefulness. It is possible, for each of these instances, to imagine situations in which they are relevant contributions to a conversation.

It is unlikely that (6) would be considered an informative utterance[2] since it states easily inferable information about a library and could be classed as overinformative or informationally redundant. One way that a listener might reconcile a speaker choosing to produce (6) might be to draw an atypicality inference: The speaker is cooperative and tries to provide newsworthy information; mentioning the presence of books at a library isn't normally newsworthy (books are an inherent property of libraries), but perhaps this library temporarily lacked books and the update about the new-found presence of books is worthy of note (e.g. Kravtchenko & Demberg, 2022).

The utterance in (7) is informative in that it provides useful information about the library; it presents information that is often sought and is actionable (i.e. opening times to allow for visiting the library) and reduces a listener's uncertainty about the world.

In (8) however, the informativity and relevance of the utterance is less immediately obvious. While it does reduce uncertainty about the world by presenting information about the library, this information is not particularly useful in that there is no clear action that a listener can take after hearing this. Similarly, the information doesn't necessarily induce an atypicality inference; blue walls aren't an inherent property of libraries, thus the utterance of (8) fails to be specifically overinformative and redundant.

[2] Although the case could be made that these utterances reveal something about the speaker themselves.

Utterances such as (8) present listeners with a choice; they could accept this as a piece of information about the world, a transparent mapping between utterance and meaning. However, the fact that a speaker chose to produce (8) at all may invite a listener to search for possible enrichments: *why* is such a trivial utterance relevant, what are the speaker's goals, and potentially what additional meaning can be inferred that would make this a relevant contribution.

In a case like (9) below, additional meaning can arise if the addressee believes the speaker is: (i) sufficiently cooperative to adhere to the general goal of conveying sufficient information for the current exchange and (ii) knowledgeable of the situation over time. If these assumptions hold, one way of reconciling the production of an utterance about a fairly trivial situation is to infer (9a), whereby what is being communicated by the utterance is not just the colour of the walls but the (rarer) event of them having been painted a new colour. In other words, (9) permits the possible inference of (9b).

(9) The library walls are blue.
    a. → *the situation has changed*.
    b. → *the walls used to be different*.

It is important to note that we are not claiming that the *something has changed* inference is a particular category of inference. Rather, it is an illustration of a potential reading that could arise from the speaker's choice to produce an utterance when the alternative was to remain silent. Note that an addressee may evaluate the utterance differently depending on their assessment of the content, the speaker, and the communicative context: In some cases, they may not retrieve any additional meaning and rather accept the information that the library walls are blue as simply the speaker's presentation of a fact about the world; alternatively, they may retrieve another inference, for example that the speaker thinks painting library walls blue is unusual; lastly, they may take the utterance as a social gesture of small talk, without much transfer of contentful information.

Speaker specificity

Previous research shows that conversational partners engage in a wide range of reasoning about each other, guiding both speakers' production decisions (e.g. Brown-Schmidt & Konopka, 2011; Lockridge & Brennan, 2002; Nadig & Sedivy, 2002; Jara-Ettinger & Rubio-Fernandez, 2021; Yoon, Koh, & Brown-Schmidt, 2012) and the meanings derived by listeners (e.g. Cai et al., 2017; Davies et al., 2022; Grodner & Sedivy, 2011; Loy, Rohde & Corley, 2019; Regel, Coulson, & Gunter, 2009; Yoon, Jin, Brown-Schmidt, & Fisher, 2021). The decision to draw an inference can be influenced by a variety of speaker-specific factors such as speaker reliability (e.g. Diesendruck, Carmel and Markson, 2010; Grodner & Sedivy, 2011; Sobel, Sedivy, Buchanan, & Henessy, 2012) or speaker fluency (e.g. Arnold et al., 2004; 2007; Loy, Rohde & Corley, 2019; Orena & White, 2015; Yoon, Jin, Brown-Schmidt, & Fisher, 2021).

One crucial factor for inferencing is speaker knowledge. In a standard model of pragmatic inferencing, information about the speaker's knowledge state is used in the derivation process (Geurts, 2010). That is, in order for a listener's inference of a speaker's intended meaning to be licensed, the listener needs to assume that the speaker has knowledge (or an opinion) about the proposition of the utterance (e.g. Schulz & Rooij, 2006; Spector, 2007). Indeed, empirical evidence supports this view as inferencing occurs more reliably when speakers are deemed knowledgeable (Bergen & Grodner, 2012; Moty & Rhodes, 2021). Bergen and Grodner (2012)

found that participants were less likely to compute a scalar implicature when the speaker had "skimmed" the document they were mentioning compared to a speaker who had "read meticulously". Moty and Rhodes (2021) showed that the inferences drawn from generic utterances can be influenced by speaker knowledgeability; inferences about group characteristics only arose when the speaker was knowledgeable. More recently, work from Zhang and Wu (2023) demonstrates that both native and non-native speakers are sensitive to speaker knowledge states and are more likely to draw inferences from knowledgeable speakers.

In the present work we ask how linguistic and extralinguistic speaker-specific characteristics contribute to the interpretation of trivial utterances and the potential inferences that arise. As noted, one challenge faced by comprehenders is recognising when to draw an inference since a key characteristic of pragmatic inferences is they are not strictly part of what is said. Although a speaker is never on record as having asserted the additional enriched content, an addressee may nonetheless be inclined to identify and infer such meaning because doing so may be necessary for maintaining the sense and relevance of a conversational contribution (Sperber & Wilson, 1995; Carston, 2004). However, the felicity of such an inference may be dependent on S*peaker Knowledge* or *Speaker Style*.

In all experiments reported here, we manipulate Speaker Knowledge by varying whether the speaker is described as being familiar or unfamiliar with the location they are talking about. In Experiment 1, we investigate whether the decision for a speaker to produce an utterance is sufficient for a comprehender to infer a goal or reason behind the utterance. In other words, do participants accept trivial utterances as simple statements of fact about the world or are they inclined to draw an inference about *why* the speaker chose to produce that particular utterance? We further investigate whether the rate at which participants draw such inferences is influenced by the presence/absence of a linguistic emphasis cue ("Hey, guess what"). The results show that participants draw inferences even in the absence of such a cue, suggesting that the availability of that inference does not depend on an explicit signal to pay attention to a speaker's contribution. Instead, there is a more general expectation that a speaker will have a reason for producing a particular utterance when they could have said nothing (see assumption of relevance, Sperber & Wilson, 1995, Wearing, 2015).

In Experiment 2, we additionally vary the Speaker Style, including or excluding a description that the speaker "isn't normally very chatty." Such a speaker presumably has a higher threshold for what is worth mentioning, and so if they produce a seemingly trivial utterance, comprehenders may be more likely to ascribe greater meaning to the choice to produce that utterance. We demonstrate that relevance inferences are computed more often when trivial utterances are produced by a knowledgeable speaker who is characterised as quiet.

Processing and pragmatic inferences

A common question regarding pragmatic inferences relates to whether or not the process of inferencing is costly. A number of studies demonstrate that it can be costly to compute pragmatic inferences. For example, in eye-tracking studies participants show a delay in fixating on the implicature-consistent target (Huang & Snedeker, 2009); in sentence verification tasks, responses consistent with an implicature response take longer than non-implicature responses (Bott & Noveck, 2004); and ERP data demonstrates a processing cost for implicature computation (Noveck & Posada, 2003). This finding that implicatures are costly does not always hold (e.g. Breheny, Ferguson & Katsos 2013; Grodner, Klein, Carbary & Tanenhaus, 2010).

Indeed, inferences are drawn with differing levels of ease, dependent on a number of constraints (e.g. type of trigger: van Tiel & Schaeken, 2017; knowledge of speaker: Bergen & Grodner, 2013; Kampa & Papafragou, 2020; Papafragou, Friedberg, & Cohen, 2018; contextual support: Breheny, Katsos, & Williams, 2005; Degen & Tanenhaus, 2015; Huang & Snedeker, 2018; Singh, 2019; van Tiel, Pankratz, & Sun, 2019). The standard account for why implicatures may be costly is that a comprehender is posited to interpret the literal meaning first (e.g. Grice, 1975; Geurts, 2010), at which point they notice the violation of their expectations in the speaker's utterance and look for a way to enrich the meaning that would allow the utterance to constitute a cooperative contribution to the conversation. This process of identifying the seeming violation, rejecting the literal meaning, reasoning about the speaker's intentions, and computing the inference is posited to be costly. In Relevance Theoretic terms, comparing the positive cognitive effects of the additional meaning against the effort of doing the inference calculation may itself be costly (Sperber & Wilson, 1995).

The process for computing relevance inferences may be much the same, whereby listeners who encounter a trivial utterance may seek to reconcile the mismatch between their expectations for a cooperative utterance and the trivial utterance encountered. However, it is not known whether their computation is costly or not. In Experiments 3 and 4, we test whether or not relevance inferences are costly to compute and at what point in processing these inferences arise. If there is a cost to computing relevance inferences, this cost may be visible early in processing, i.e. upon encountering a trivial utterance, or at a later stage when prompted to endorse the inferred meaning.

Experiments overview

We present 4 studies investigating relevance inferences arising from trivial utterances (e.g. "the library walls are blue"). Experiments 1 and 2 are offline studies manipulating expectations for relevant and informative conversational contributions through linguistic and extralinguistic features to probe the availability of relevance inferences. Specifically, we manipulate properties of how the speaker speaks (Experiment 1: presence or absence of an Emphasis Cue "Hey, guess what") and how they are characterised (Experiment 2: quiet or not). The emphasis cue and the quiet characterisation may both highlight the communicative intentionality of the speaker's utterance and enhance comprehenders' search for additional meaning. In both experiments, we also manipulated Speaker Knowledge by varying whether the speaker describes a location they are familiar with or a location they are unfamiliar with. If comprehenders reason about why a speaker might produce a trivial utterance, one candidate explanation is that the described situation represents a departure from the usual state of affairs. However, this reasoning only follows if the speaker is knowledgeable about a location over time.

To assess if comprehenders do indeed compute relevance inferences, we use a probe question to ask participants about what the depicted situation (e.g. the colour of the library walls) used to be like: A response that the situation used to be *different* suggests that the comprehender has inferred meaning beyond what was said. Our findings demonstrate that relevance inferences are computed regardless of emphasis cue, suggesting that the availability of these inferences does not depend on explicit linguistic cueing. Extralinguistic cues, however, are shown to affect the availability of relevance inferences; these inferences occurred more often when trivial utterances were produced by a knowledgeable speaker and one who is

characterised as quiet, suggesting that the availability of these inferences increases when the comprehender posits that there must be an identifiable reason for the form used.

In Experiments 3 and 4, we test the processing of relevance inferences using a sentence verification paradigm. We ask how readily relevance inferences are derived; are they computed as a part of normal comprehension processes? In other words, as soon as a comprehender encounters a trivial utterance, is the mismatch between their expectations and the reality of the trivial utterance reconciled via the calculation of an inference, and is this inference costly? Or do these inferences arise only if a comprehender is specifically prompted to consider the speaker's goals? If relevance inferences arise immediately upon, or very soon after, encountering a trivial utterance, the inference computation would be predicted to incur an immediate cost, observable in longer reading times when processing the utterance itself. If these inferences do not arise automatically, there may be a delay only at a later point when the comprehender is specifically prompted to compute the inference. A third possibility is that relevance inferences are so common as to not carry any cost. Based on the reading time data and response times, there appear to be processing costs associated with computing such an inference and the cost emerges at the point when participants are prompted to compute an inference.

Data availability

All studies reported were carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki). We obtained ethical approval from [blind for peer review] research ethics committee and all participants gave informed consent. In line with open science practices all materials, data, and analysis scripts are available on the open science framework (https://osf.io/nbhya/overview?view_only=fa86cfb308564469ac6f7170d7fe03b9).

Experiment 1

Experiment 1 used a 2x2 design to manipulate a linguistic Emphasis Cue (presence/absence of "Hey, guess what") and the extralinguistic factor of Speaker Knowledge (familiarity/unfamiliarity with the location). The expression "Hey guess what" may act as an explicit linguistic signal to highlight that the upcoming content is newsworthy and trigger an expectation that the speaker's contribution is going to be relevant (Bates, 1974; Dubois, 1989), thereby allowing us to test whether relevance inferences are readily available or only arise in contexts in which a speaker emphasises their intention to communicate.

Speaker Knowledge was manipulated by varying whether or not the speaker was knowledgeable about the location they were talking about. If interpretation depends on the reasoning outlined above, whereby familiarity with a location is central to drawing the specific relevance inference that the speaker is describing a situation that used to be different, the familiar condition should yield greater rates of inferencing than the unfamiliar condition. Additionally, if an explicit cue to the newsworthiness of the upcoming content increases the availability of such relevance inferences, these rates should be boosted in the cued condition; alternatively, if expectations for cooperativity are pervasive, the potential inference may be drawn regardless of any cues.

Method

Participants

As determined by an a priori power analysis based on pilot data (available on osf), data were collected online via the crowdsourcing site Prolific (prolific.ac) from 200 participants. Due to a technical error, data were collected from an additional 3 participants resulting in 203

participants (Age 18-64, *M*= 28.11years). All participants were fluent speakers of English and received payment at a rate of £8.14 per hour for their participation.

Design and materials

The Emphasis Cue "Hey, guess what" was manipulated between participants; half the participants saw the uncued form of the utterance (Figure 1) and the other half saw the cued form of the utterance (Figure 2). Speaker Knowledge was manipulated within participants in a blocked design; one block for each of two locations. Participants were presented with a speaker, Suzy, who is telling her dad about her day in a familiar (School) or unfamiliar location (Prime Minister's office; PM), with the order of blocks counterbalanced across participants. Participants are told, for the familiar condition, that Suzy has gone back to School after the summer holidays and, in the unfamiliar condition, that she has been on a School trip to the Prime Minister's office.

In total there were 20 critical utterances of the form "I saw that…" which described plausibly changeable situations: e.g., "I saw that the library walls are blue", "I saw that there are mirrors in the hallway", and "I saw that the chairs have cushions" (for a full list see supplementary materials). Since walls can be painted, mirrors can be hung on walls, and cushions can be placed on chairs, it is plausible that these situations may have changed with relative ease. We chose situations such as these rather than situations that are constant (e.g. white lines at the zebra crossing; since white lines are an essential feature of a zebra crossing) or situations that must inevitably change (e.g. puddles on the ground, since puddles form and evaporate based on the weather). Participants saw 10 items in the familiar condition and 10 in the unfamiliar condition. Across participants, all utterances were seen in all conditions, but each participant only saw a given item once. Utterances were presented visually in a speech bubble alongside a picture of Suzy (Figs. 1 & 2).

For each item, participants were asked "What do you think it was like a few months ago? Same or Different?" We also included two attention-check control items about situations that are likely to change over time ("I saw my friend graze their knee" and "I saw that the leaves on the tree are turning orange"), for a total of 22 items.

To check that the Speaker Knowledge manipulation was understood as intended, we included a Familiarity Task in which participants were asked to judge how familiar they thought Suzy and her dad were with the two locations. Responses were elicited on a scale of 1-5, with 1 being not familiar at all and 5 being very familiar. It was assumed that participants would accept that Suzy and her dad were more familiar with School than with the Prime Minister's office.

Figure 1. Experiment 1 stimuli item with emphasis cue absent

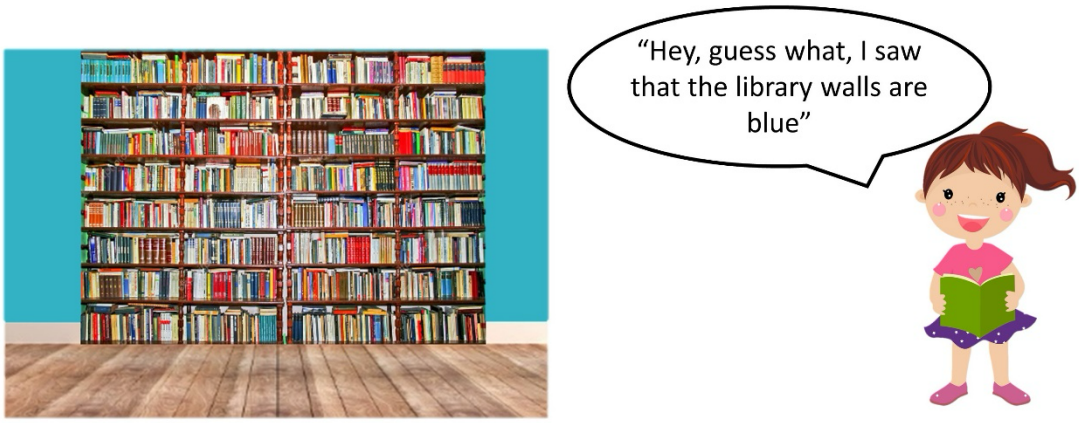

Figure 2. Experiment 1 stimuli item with "Hey, guess what" emphasis cue present

Procedure

The study was hosted and administered online through Qualtrics surveys (Qualtrics.com). Item presentation was blocked and counterbalanced to ensure an equal number of participants saw the familiar versus unfamiliar block first and trials within each block were fully randomised. At the start of a block, participants were introduced to Suzy, her dad, and the location they were going to be talking about. On each trial the critical items were presented (as in Figures 1 & 2) along with a prompt question on the same screen. Participants were asked "What do you think it was like a few months ago?" and could respond either "same" or "different". Responding "different" indicates (or is at least compatible with) participants' acceptance of an inference, namely that the speaker's utterance conveys additional meaning that something has changed. Following the main task, participants completed the Familiarity Task. Finally, participants were asked three optional feedback questions about the study: "What do you think this study is about? Did you use a strategy when answering? Do you have any other feedback about the experiment?"

Results

To ensure we only analysed data from participants who were concentrating on the task, we excluded data from 6 participants who failed both attention check questions. The remaining 197 participants answered at least one attention question correctly and were included in the analysis (97 participants in the Emphasis Cue, "HeyGuess", group and 100 in the uncued group).

To assess the success of the Speaker Knowledge manipulation, the ratings data were analysed using Friedman's test, and pairwise comparisons were run with Wilcoxon signed rank test (see supplementary materials). Participants' familiarity ratings confirmed the Speaker Knowledge manipulation: Participants rated Suzy and her dad as being more familiar with School (see Table 1) than the Prime Minister's office and they rated Suzy as being more familiar with School than her dad (χ^2 (3)=367.2, p<.001). Suzy and her dad were rated as being equally unfamiliar with the Prime Minister's office.

Character	Location	Mean rating	Standard deviation
Suzy	School	4.01	.91
Suzy	PM	1.80	1.80
Dad	School	2.96	.94
Dad	PM	1.83	1.02

Table 1. Mean familiarity ratings for Experiment 1

For the primary analysis, we modelled the binary responses (no inference "same"/ inference-endorsing "different") with a logistic regression in R (Version 4.0.3, R core team, 2020) using lme4 (Version 1.1-23; Bates, Mächler, Bolker, & Walker, 2015) and emmeans (Version 1.6.0; Lenth et al. 2021) with fixed effects for Speaker Knowledge, Emphasis Cue, and their interaction.

For all models reported throughout the paper, we started with maximal random effects, consistent with the guidelines for model building for psycholinguistic hypothesis testing as specified in Barr et al. (2013). In cases of non-convergence, we used the maximal model that allowed for convergence by iteratively simplifying the random effects' structure (see supplementary materials for models).

Variables were centred such that, for Speaker Knowledge, the unfamiliar PM office was coded as -0.5 and the familiar School as 0.5. For the Emphasis Cue, uncued was coded -0.5 and cued "HeyGuess" was coded 0.5.

Figure 3 shows the proportion of responses in which participants endorsed the inference by indicating that the location was different a few months ago. As predicted, we find a main effect of Speaker Knowledge (β = -1.43, SE=.192, z =-5.96 p < .001) with participants endorsing the inference more often for the familiar location (School) than the unfamiliar location (PM). The effect of the Emphasis Cue was marginally significant in the predicted direction with more endorsement of the inference when the utterance contained an Emphasis Cue (β= -.32, SE=.185, z = -1.74 p = .083). There was no interaction (β=.113, SE=.314, z = .036 p = .971). The Speaker Knowledge finding is in line with the claim that inference rates reflect comprehenders' perception of the speaker's extralinguistic knowledge state: Participants were more likely to infer that something had changed when the speaker was familiar with the location they were talking about. Participants also relied on their own world knowledge when responding. In the post-study feedback (see supplementary materials), some participants indicated that they considered the inferred changes to have been too effortful/costly to be likely; "*the only thing I kept in mind was that quite large things were more likely to have been there for a long time (like picnic benches, a shed or a sculpture) but smaller things may have changed over time (like the soap or mirrors on a wall)"*.

The lack of an effect or interaction with Emphasis Cue can be taken to show that inferencing rates do not depend on the speaker's emphasis on newsworthy content, suggesting that the derivation of such inferences can arise without an explicit linguistic cue to compute additional

meaning. Given the role played by the speaker's knowledge state, an extralinguistic factor, Experiment 2 considers another property of the speaker – their speech style.

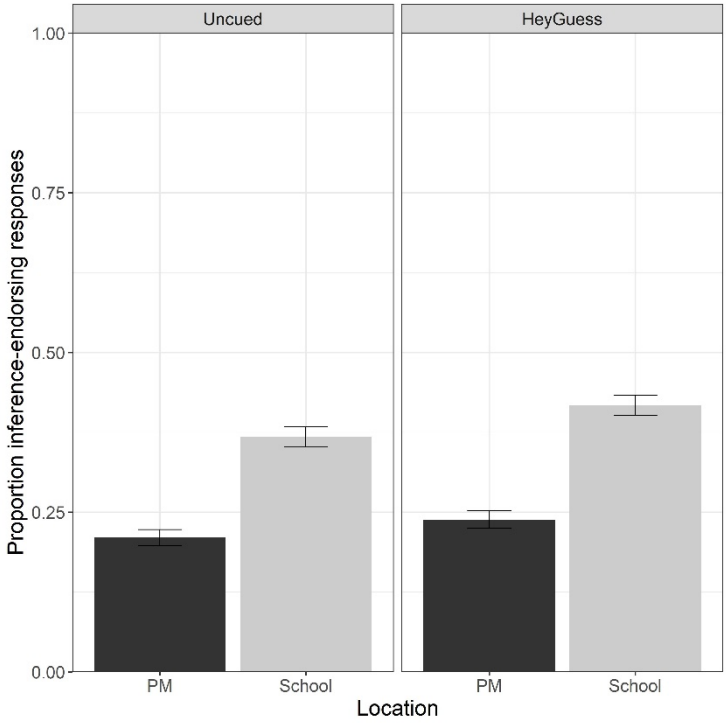

Figure 3. Proportion of inference-endorsing responses by Speaker Knowledge (familiar School vs unfamiliar Prime Minister's office) and Emphasis Cue (Uncued vs "Hey Guess What") from Experiment 1. Inference-endorsing corresponds to a "different" response to the question "What do you think it was like a few months ago?"

Experiment 2

Experiment 1 showed that relevance inferences are not dependent on explicit cues. Rather, the computation of these inferences hinges on the perceived Knowledge of the speaker. When the speaker was judged to be knowledgeable, participants were more likely to compute relevance inferences than when the speaker was deemed not knowledgeable. However, even when the speaker was not knowledgeable, participants were still inclined to search for additional meaning.

The goal of Experiment 2 is twofold. Firstly, we look to replicate the Speaker Knowledge finding. Secondly, we investigate further extralinguistic speaker-specific factors that may influence the computation of relevance inferences, specifically *Speaker Style*. In one condition, the speaker is characterised as "quiet" and "not very chatty" and in the other condition no information about the speaker was provided. Knowing that a speaker is not very talkative may impose a higher threshold on what that speaker will deem an appropriate conversational contribution. Consequently, encountering a quiet speaker who produces a trivial utterance may prompt greater rates of inferencing due to the rarity of their speaking at all and therefore increased the possibility that the speaker is using this single utterance to convey additional meaning. Such a finding would underscore the role of speaker characteristics in models of cooperative communication.

Method

Participants

200 participants, who reported being fluent in English, were again recruited from Prolific. Due to a technical error, data were collected from an additional 5 participants resulting in a total of 205 participants (Age 18-55, *M*= 28.33 years). All participants received payment at a rate of £8.14 per hour for their participation.

Materials & Design

As in Experiment 1 we manipulated Speaker Knowledge. We also manipulated Speaker Style between participants: Half of the participants were told that "Suzy isn't normally very chatty. Her family and teachers think she is quiet." The other half were told nothing about the speaker. We used the same familiar/unfamiliar location materials as in Experiment 1, all in the uncued form. Alongside the 20 target items, again there were two attention-check controls, for a total of 22 items.

Procedure

Procedure was the same as Experiment 1.

Results

The same analysis procedure was used as in Experiment 1. There were 13 participants who failed the two attention check questions and were removed from the analysis. The remaining 192 participants answered at least one attention question correctly and were included in the analysis (94 participants in the "quiet Suzy" group, 98 in the "normal Suzy" group). Again, we assessed the success of the Knowledge manipulation (see supplementary materials). As in Experiment 1, participants rated Suzy and her dad as being more familiar with School (see Table 2) than the Prime Minister's office and they rated Suzy as being more familiar with School than her dad ($\chi^2(3) = 359.1$, $p<.001$). Suzy and her dad were rated as being equally unfamiliar with the Prime Minister's office.

Character	Location	Mean rating	Standard deviation
Suzy	School	4.04	.86
Suzy	PM	1.96	1.05
Dad	School	3.03	.76
Dad	PM	1.96	.94

Table 2. Familiarity ratings for Experiment 2

For the analysis, we used a logistic regression with fixed effects for Speaker Knowledge, Speaker Style, and their interaction, as well as the maximal random effects structure that allowed model convergence. Variables were centred such that for Speaker Knowledge, again the unfamiliar PM office was coded as -0.5 and the familiar School as 0.5. For Speaker Style, *Normal* was coded -0.5 and *Quiet* was coded 0.5. Figure 4 shows the proportion of responses in which participants said the location was different a few months ago.

The effect of Speaker Knowledge found in Experiment 1 was replicated here with greater rates of inferencing for the Familiar (School) than Unfamiliar (PM) location (β= 1.37, SE=.178, z = 7.70 p < .001). As predicted, the results also showed an effect of Speaker Style with greater rates of inference endorsement for the Quiet Speaker condition (β= .87, SE=.274, z = 3.41 p = .001). There was also an interaction whereby the effect of Speaker Knowledge was larger in the Normal Speaker condition (β=-.76, SE=.341, z = -2.23 p = .026). Follow-up analyses confirmed

that the Speaker Knowledge effect was present and in the same direction for both the Normal (z=6.74, $p < .001$) and Quiet Speaker (z=4.25, $p = .001$).

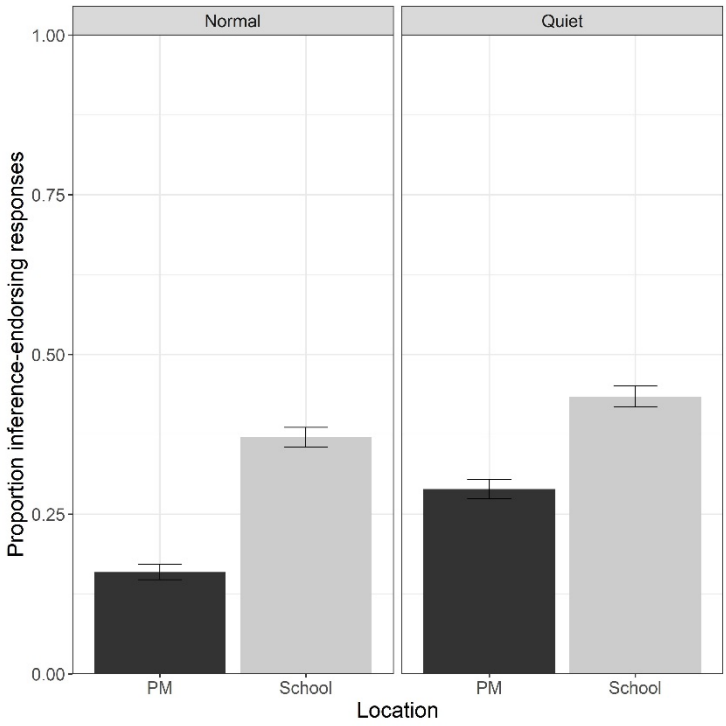

Figure 4. Proportion of inference-endorsing responses by Speaker Knowledge (familiar School vs unfamiliar Prime Minister's office) and Speaker Style (Normal vs Quiet) from Experiment 2

Interim discussion

Experiments 1 and 2 demonstrate that, as with other forms of inferencing, speaker-specific factors play a significant role in the computation of relevance inferences. With a speaker deemed knowledgeable of a situation over time, participants were more likely to respond that the described situation used to be different, suggesting they were drawing inferences from trivial utterances in order to reconcile their expectations for cooperative communication. Similarly, when a speaker who was described as quiet produced trivial utterances, participants were more likely to interpret the utterance non-transparently as conveying additional information. We posit that the increased rates of inferencing are due to comprehenders judging that speakers who are quiet have a higher threshold for what constitutes an acceptable conversational contribution. In other words, hearing that a quiet speaker has chosen to utter anything at all is signal enough that they are intending to communicate something interesting.

Experiment 1 found that providing a cue that emphasized newsworthiness did not significantly increase inference rates. This finding is perhaps not surprising if we assume that, by the very nature of participating in a conversation and speaking to another person, a speaker is already committing to the cooperative exchange of information. It is likely that as a default, speakers are expected to be cooperative and this expectation changes only when we have evidence of a speaker's non-cooperativity. Additionally, Experiment 2 demonstrated that comprehenders adjust their expectations based on speaker-specific characteristics. Upon encountering a quiet speaker, participants were more likely to derive relevance inferences.

The results also showed an unexpected interaction whereby the difference in inferencing rates by location (knowledge) was larger for the normal speaker than for the quiet speaker. This difference was driven by an increase in inferencing in the unfamiliar quiet speaker condition. Deriving a "something has changed" inference in the unfamiliar condition is infelicitous since the speaker is unlikely to have knowledge of the situation beforehand and thus, they are not in a position to be conveying additional information about a change of state. We speculate that the interaction may reflect a determination on the participants' part that a quiet speaker, who makes trivial comments about an unfamiliar location, must be trying to convey *some* additional meaning (why would Suzy mention the colour of the walls, or the presence of mirrors, or cushions on the chairs at the Prime Minister's office? What else is she trying to communicate?), but the only way to convey the derivation of an inference of *some* additional meaning given the constraints of the task was to respond "different". Of course, during normal interaction, we might expect the speaker to elaborate on the things they've said to make their motivations clearer, or as an addressee we could probe the speaker for more information. In the present task, however, this was not possible. A future study could probe the nature of such inferences, possibly with a more open-ended task (as in Ryzhova, Mayn, & Demberg, 2023).

Experiment 3

We now turn to the questions of how readily relevance inferences are derived and whether they incur a cost in a full-sentence reading task and a sentence verification task (as in Bott & Noveck, 2004 and Shukla, et al., 2022). Sentence verification tasks involve presenting statements for participants to judge as true or false. Prior work has used sentence verification tasks to test underinformative utterances containing a scalar trigger, such as "Some elephants have trunks", which allows for two possible interpretations: These statements are true under a non-implicature interpretation (*some and possibly all elephants have trunks*) and false if an inference has been computed (*some but not all elephants have trunks*). Bott and Noveck (2004) found that participants who responded "false" to underinformative sentences were slower to respond than participants who responded "true". Bott and Noveck took this as evidence of a cost for deriving scalar implicatures (see also De Neys & Schaeken, 2007).

In addition to measuring rates of inferencing via sentence verification, we assess two measures of processing. First, we measure the time taken to read the utterance itself. Second, we measure the time taken to respond to the sentence verification question "was it the same a few months ago?" with a yes/no button press. Responding "No" to this prompt question is consistent with drawing a relevance inference (*No, it was not the same a few months ago*) whereas "Yes" responses indicate that no inference has been drawn (*Yes, it was the same a few months ago*). This response pattern ("No" as the inference response) is analogous to that of previous sentence verification tasks that have probed the costs associated with computing scalar implicatures (Bott & Noveck, 2004; Shukla et al., 2021; van Tiel, Pankratz & Sun, 2019; see also Stivers et al., 2009).

If relevance inferences are computed as a part of normal comprehension processes—i.e. upon encountering a trivial utterance—and if their computation incurs a cost, this cost should be visible immediately, observable in longer reading times when processing the utterance itself (particularly in a felicitous inference context such as the School condition). However, relevance inferences may not arise automatically, in which case there may be a delay to seeing the cost, i.e. at a later point when the inference is specifically prompted and an inference-endorsing response is given. There is a third possibility that relevance inferences may not be costly to

compute at all, in which case we would not see any differences in reading times or response times.

Method

Participants

We recruited 200 participants, who reported being fluent in English, from Prolific. Two participants failed to complete the experiment and were removed from all analyses. Participants were paid at a rate of £8.26 an hour.

Design & Materials

We manipulated Speaker Knowledge with a familiar and unfamiliar location as before. The materials included the same 20 target items used in earlier experiments (10 per location), with no emphasis cue. In this experiment, there were 10 filler items and one attention check per location for a total of 42 items (21 items per location, see supplementary materials for full list).

Procedure

The experiment was hosted and administered online through Psytoolkit (Stoet, 2010; 2017). As in Experiments 1 and 2, item presentation was blocked and counterbalanced to ensure an equal number of participants saw the familiar versus unfamiliar location first and trials within each block were fully randomised. At the start of a block, participants were introduced to Suzy, her dad, and the location they were going to be talking about. They were then presented with the sentences. Following the presentation of a sentence in critical and attention check trials, on a separate screen, participants were asked the prompt question "Was it the same a few months ago?" and could respond either "Yes" or "No". Responding "No" is compatible with participants' derivation of additional meaning, whereas responding "Yes" is consistent with a transparent interpretation of the utterance. Filler trials were presented without a subsequent question screen.

Reading times were calculated as the time spent on the sentence presentation screen where the utterance appeared. Response times were measured as participants' time to respond to the question on the subsequent screen. All figures present the untransformed times, and all statistical testing was carried out on the log-transformed times.

Results

A total of 194 participants' data were included in the analysis; 4 participants were excluded for failing both attention check questions. All analyses were carried out in R (Version 4.0.3, R core team, 2020) using lme4 (Version 1.1-23; Bates, Mächler, Bolker, & Walker, 2015) and emmeans (Version 1.6.0; Lenth et al., 2021). We always used the maximal model that allowed for convergence (see supplementary materials). We analysed the binary responses (no inference "Yes" / inference-endorsing "No") using a logistic regression with Speaker Knowledge as the predictor. We analysed the log-transformed response times and reading times for critical sentences using regression models with Speaker Knowledge, Response Type, and their interaction as predictors. Variables were coded such that, for Speaker Knowledge, the unfamiliar PM office was coded as -0.5 and the familiar School as +.05 and for Response Type, non-inference responses ("Yes") were coded as -0.5 and inference-endorsing responses ("No") were coded as +0.5.

The reading time and response time data were cleaned prior to log-transformation such that reading times shorter than 500ms, response times shorter than 200ms, and any times that were

more than two standard deviations above the mean were excluded. This accounted for 293 trials, leaving 3764 trials for analysis.

Reading times

We measured the time participants spent reading the critical utterances before pressing a button to continue to the question prompt. Figure 5 shows the untransformed reading times on the critical utterances by location and response type (i.e. the response participants went on to give to the probe question). As Figure 5 shows, the longest reading times corresponded to the circumstances in which participants were reading about the familiar location condition (the one which yielded more inferences) and subsequently indicated that they had drawn an inference ("No" response). These untransformed means[3] hint at a numeric pattern whereby inferencing induces a cost during processing. However, when analysing the log-transformed reading times none of the factors or interactions reached significance (Speaker Knowledge: β=1.945, SE=.013 t=1.454, p=.146; Response type: β=-.006, SE=.016, t=-.390, p=.698; Speaker Knowledge X Response interaction: β=-.006, SE=.028, t=-.205, p=.837). Follow-up Bayesian analysis (Morey & Rouder, 2024) indicates that there is insufficient power to detect differences in the reading times based on the individual factors; however, they do indicate that there is no interaction (Speaker Knowledge BF = .006, Response Type BF = 0.122, Speaker Knowledge X Response interaction BF = .292).

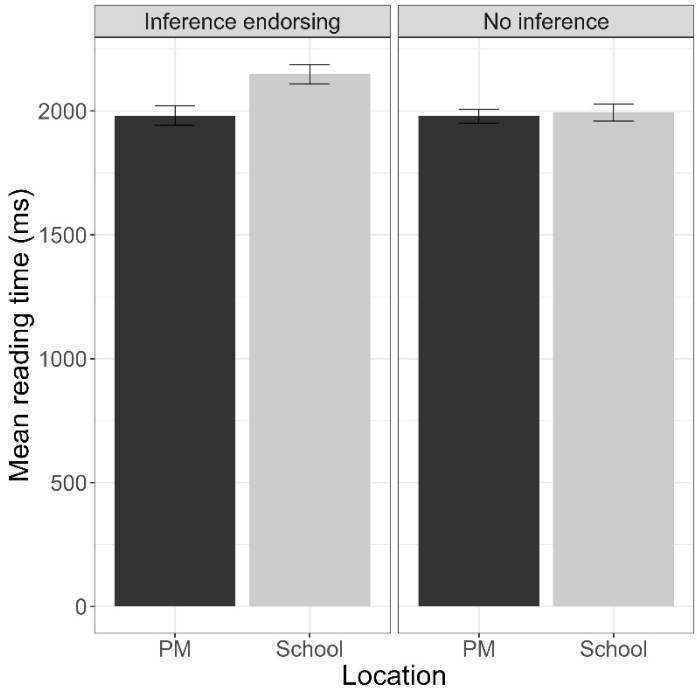

Figure 5. Mean utterance reading times from Experiment 3 by location and subsequent inference response

Inference rates

As seen in Figure 6, the effect of Speaker Knowledge found in Experiments 1 and 2 was replicated here with a greater rate of inference-endorsing responses ("No" *it was not the same*) for the familiar (School) location than the unfamiliar (PM) location (β=.976, SE=.083, z=-11.743 $p < .001$).

[3] Untransformed mean reading times for Inference Endorsing in School 2149ms, No inference in School 1993ms

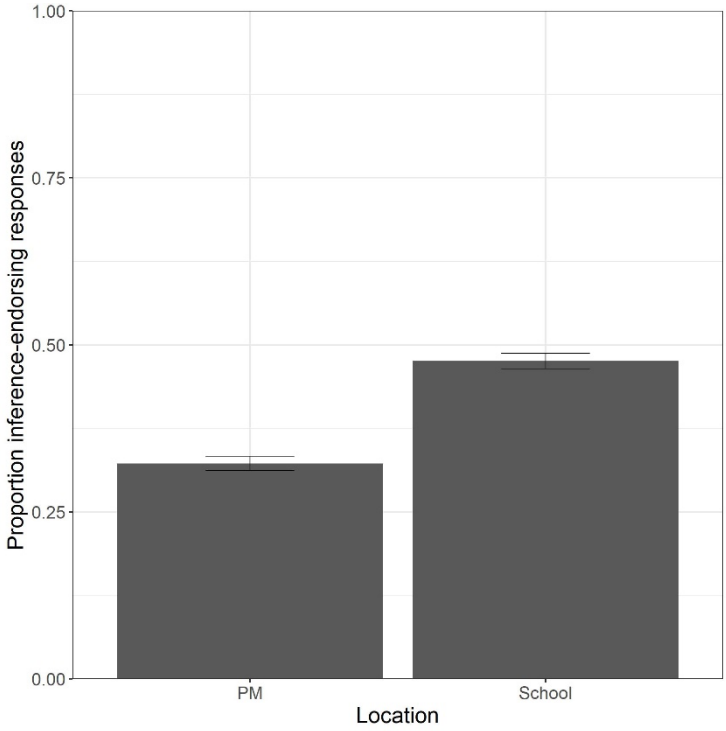

Figure 6. Proportion of inference-endorsing responses, from Experiment 3, by Speaker Knowledge (familiar School vs unfamiliar Prime Minister's office); in response to the question "was it the same?", "No" represents the inference-endorsing response.

Response time

Along with measuring participants' binary responses, we also measured the time taken to respond. If inferences are costly to compute, such computation (corresponding to the "No" response) should yield slower reading times. Thus, we included Response Type as a fixed effect in the model and tested the interaction with Speaker Knowledge. However, responding "No" is generally slower than responding "Yes" (e.g. Bott & Noveck, 2004; Carpenter & Just, 1975; Wason, 1961; Wang, et al., 2021), a point we come back to in Experiment 4.

As shown in Figure 7, we see a main effect of Response Type whereby inference responses ("No") are indeed slower than no-inference responses ("Yes") (β=.172, SE =.020, t=8.540, p<.001). There is no significant effect of Speaker Knowledge (β=.016, SE =.017, t=.893, p=.372).

However, there is an interaction between Speaker Knowledge and Response Type (β=-.077, SE=.037, t=-2.077, p=.038). The direction of the interaction is not specifically useful for establishing whether there is a cost to drawing an inference when Speaker Knowledge (School location) supports the inference. Rather, the slowdown for inference-endorsing responses ("No" responses) is larger in the non-inference-supporting condition (PM) than in the inference-supporting condition (School). While inference responses in the PM location are infelicitous (since it is highly unlikely Suzy knew what the PM's office was like before the trip), this slowdown mirrors our earlier finding in Experiment 2 where we saw an increase in inferencing in the unfamiliar location for the quiet speaker.

Follow up analyses of the simple effects and Bayesian analysis show that if we look just at these inference-endorsing responses, such responses do not vary significantly by Speaker

Knowledge (β=.023, SE=.028, z=.797, p=.856, BF = .067), whereas the no-inference responses show a marginal difference by Speaker Knowledge with participants being slower to indicate their no-inference response in the inference-supporting (School) condition (β=-.054, SE=.023, z=-2.404, p=.076, BF=.511).

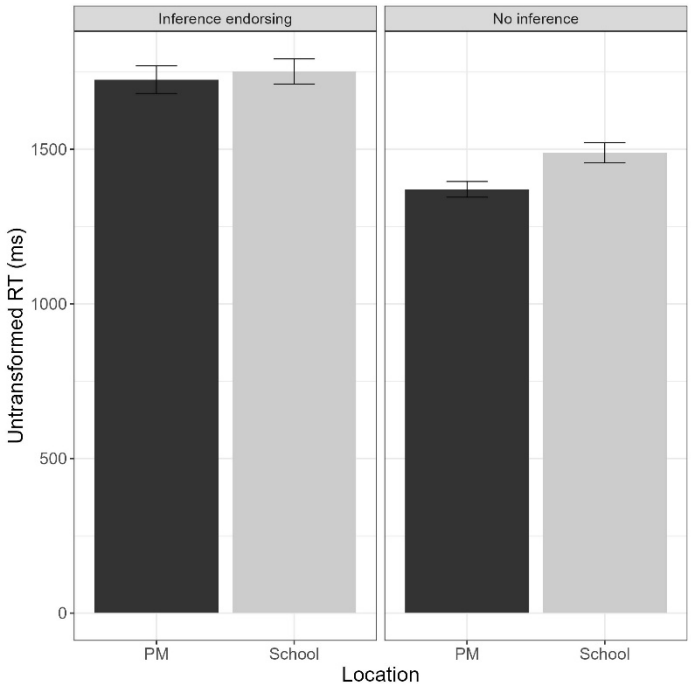

Figure 7. Response times from Experiment 3 by response and location.

Discussion

Experiment 3 replicated the findings from Experiments 1 and 2, in which there were greater rates of inference derivation when the speaker was knowledgeable.

Despite a numerical difference in reading times, the statistical analysis does not provide evidence that relevance inferences are derived during the sentence itself; rather we only see evidence of inference computation in our offline measure when participants are prompted to answer a question about the potential inference ("Was it the same a few months ago?").

Response times were slowest when participants provided an inference-endorsing response. However, there was also an interaction whereby we see a larger difference in response times when participants do not endorse the inference; participants were slower to respond "yes" in the familiar (School) condition. However, participants took longer to give a "no" response irrespective of the Speaker Knowledge manipulation. It is unclear, however, if this delay can be attributed to the derivation of a relevance inference (i.e. *No, it was not the same*) or if the increase in response time is a consequence of the general finding that it takes longer for participants to give a negative response (cf. Bott & Noveck, 2004; Stivers, 2009; Wang et al, 2021).

Thus, to address this confound, in Experiment 4 we replicate Experiment 3 but change the wording of the question from "Was it the same a few months ago?" to "Was it different a few months ago?". Thus, it is now the "Yes" response which is indicative of an inference. If relevance inferences are costly to compute, then it is expected that it will take longer for

participants to provide an inference-endorsing response “Yes” than a non-inference response “No”.

As we will show, participants in Experiment 4 do not show longer response times for “Yes” than “No”, contra the prediction if inference-endorsing responses are costly. To address the cost of relevance inferences, we conduct a cross-experiment comparison of response times based on Response Type (yes/no). We will show that across Experiments 3 and 4, we do not find a significant difference in “No” response times regardless of whether this corresponds to an inference or not. There is, however, a marked difference across experiments in “Yes” response times. When “Yes” responses correspond to an inference, participants take significantly longer to respond than when “Yes” does not correspond to an inference.

Experiment 4

Participants

A total of 213 participants were recruited from Prolific (N=110) and the student population at [University] who received course credit (N=103). All participants were fluent in English and the Prolific participants were paid at a rate of £9.10.

Method

The same materials as in Experiment 3 were used but we changed the polarity of the question to ask, “Was it different a few months ago?”. Thus, Experiment 3 and Experiment 4 differ in the inference-endorsing response: Here in Experiment 4, a “Yes” response now corresponds to an inference-endorsing response and “No” corresponds to a non-inference response.

Results

Data exclusion followed Experiment 3. There were 19 participants who failed the two attention check questions leaving 194 participants for the analysis. Following the same procedure as before for reading and response times, 401 trials were excluded leaving 3659 trials for analysis. The statistical modelling procedure followed Experiment 3 with factors for Speaker Knowledge (-0.5 for PM, +0.5 for School), Response Type (-0.5 for no inference “No” responses, +0.5 for inference-endorsing “Yes” responses), and their interaction.

Reading times

As in Experiment 3, sentence reading times show no main effect of Speaker Knowledge (β=.020, t=1.304, SE = .015, p=.193, BF = .04) nor of Response Type (β=-.015, t=-.857, SE = .017, p=.392 = .05, BF=.05) and the follow up Bayesian analysis indicates anecdotal evidence in support of no differences.

However, we do see an interaction between Speaker Knowledge and Response Type (β=-.0771, t=-2.414 SE= .032, p=.016) in the predicted direction to indicate processing cost for an utterance in a context that licenses an inference and for which the participant will endorse the inference. Follow-up analyses show that this interaction is driven by a marginally significant effect of Speaker Knowledge for participants who endorse the inference (“Yes” responses; β=-.059, SE=.034, z =-2.464 p=.066). Participants who responded “Yes”, endorsing the inference, had longer reading times in the inference-supporting (School) condition than participants who responded “Yes” in the non-supporting condition (PM). There is no significant difference in reading times across conditions for the “No” (no inference) response (p=.797, BF=.072) nor for reading times in the non-supporting (PM) condition (p=.113, BF=.081).

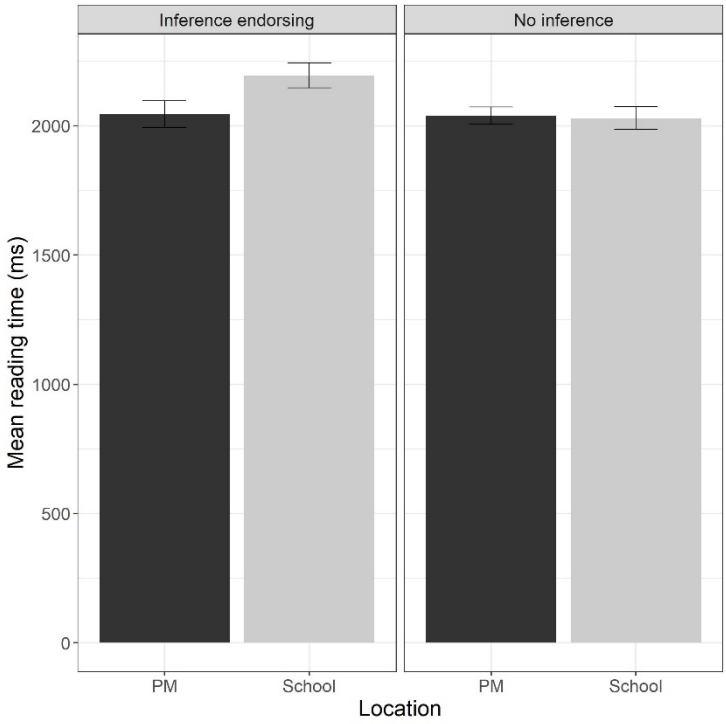

Figure 8. Reading times from Experiment 4 by Response Type and Speaker Knowledge

Inference rates

As in Experiments 1, 2, and 3, there was a greater rate of inference-endorsing responses (here, "Yes" responses) in the knowledgeable speaker condition (see Fig 9). We found a main effect of Speaker Knowledge (β=.977, SE=.079, z =12.376 p < .001); participants again derived more relevance inferences when the speaker was talking about the familiar location.

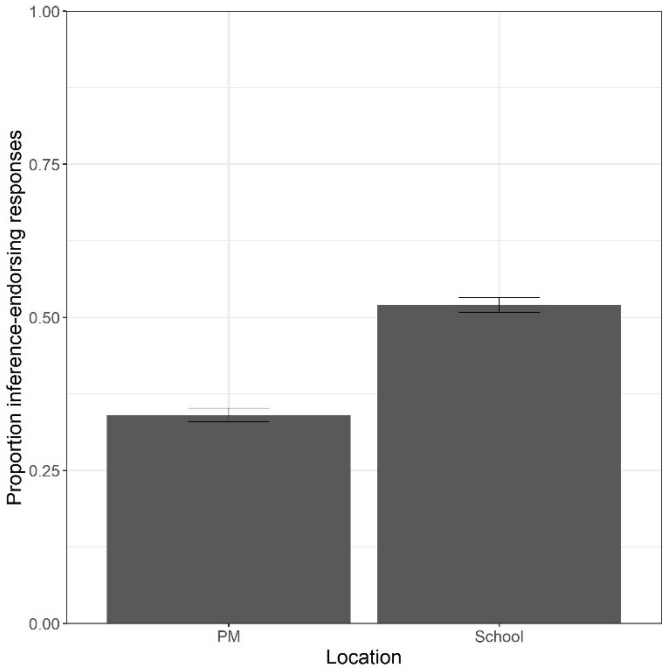

Figure 9. Proportion of "Yes" responses from Experiment 4 by Speaker Knowledge (familiar School vs unfamiliar Prime Minister's office); in response to the question "was it different?", "Yes" represents inferring additional meaning

Reaction time

The response time data shows a markedly different pattern to that of Experiment 3 (Fig.10). In Experiment 3 participants took longer to give an inference-endorsing response "No", which may have reflected either a cost for inference computation or else a general slowdown for negative answers. Here there is no significant difference in response times across Response Type (β=-.001, t= -.946, SE=.002, p = .344, BF=.202): Inference-endorsing "Yes" and no-inference "No" responses took the same amount of time. There is also no significant effect of Speaker Knowledge (β=.006, t= .313, SE=.002, p = .754, BF=.044), but there is an interaction between Speaker Knowledge and Response Type (β=.081, t= 2.092, SE=. 039, p = .037) whereby the effect of Speaker Knowledge is greater in the response times for inference-endorsing "Yes" responses. However, follow up analyses did not indicate any significant differences between conditions (p's > .148, BF's response type: No=.068, Yes= .067, BF location = .707).

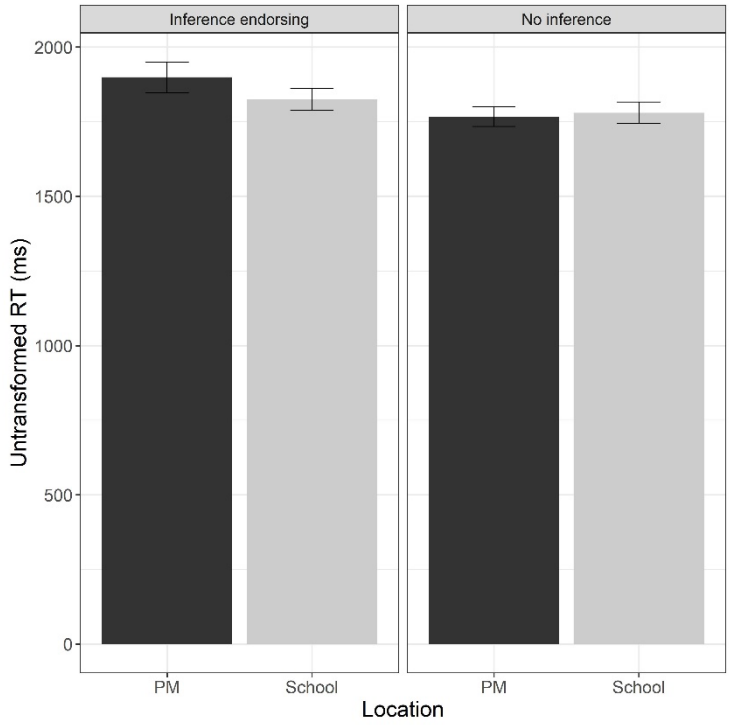

Figure 10. Experiment 4 response times by Response Type and Speaker Knowledge. Inference-endorsing responses correspond to selecting "Yes"

The results of Experiment 4 in isolation leave open the question of whether computing a relevance inference is costly or not. In Experiment 3 we saw evidence compatible with a cost to computing relevance inferences; participants were slower to respond with an inference-endorsing "No" than with no inference ("Yes). However, this may simply reflect the cost of responding "No" rather than a cost of inferencing. In Experiment 4, the sentence reading times show the predicted pattern corresponding to a cost for drawing an inference such that the longest reading times arise when an inference is licensed (the inference-supporting 'School' condition) and the participant goes on to endorse the inference (the same pattern of reading times that were found numerically in Experiment 3). Experiment 4, however, does not show a clear difference in reaction times at the inference endorsement step; there is no significant

difference in response times between responding “Yes” or “No”. Or in other words, in Experiment 3 responding “Yes” is faster than responding “No”, but in Experiment 4 there is no significant difference. This hints at there being a cost associated with relevance inferences; however, in isolation we cannot be sure since there is no significant difference between endorsing the inference and not making an inference in Experiment 4. Thus, we conducted a cross-experiment analysis of the response times.

Cross-experiment analysis

To further assess whether relevance inferences are costly to compute or not, we now compare response times from across Experiments 3 and 4. In Experiment 3, inference-endorsing responses were indicated with “No” responses, whereas in Experiment 4, inference-endorsing responses were indicated with “Yes” (boxes highlighted in Figure 11). If relevance inferences are costly, this should be most clearly seen by a slowdown in response times for “Yes” responses when this corresponds with an inference. That is, we expect “Yes” responses in Experiment 4 to be slower than “Yes” responses in Experiment 3. If the response times for “No” are guided by across-the-board slowdowns for giving a negative response, such slowdowns may conceal more subtle inference-driven differences.

We aggregated the data from Experiment 3 and 4 and compared log-transformed response times as a function of Response Type (“Yes” -.5, “No” +.5) and Experiment (Exp 3 -.5, Exp 4 +.5).

The analysis showed a main effect of Experiment (β=.255, t=6.175, SE=.0413, p<.001), a main effect of Response Type (β=.175, t= 8.730, SE=.020, p<.001), and an interaction (β=-.194, t=-6.870, SE=.028, p<.001).

In pairwise contrasts, we did not find a significant difference in how long it took to respond “No” across the studies (β=.061, z=1.473, SE=.042, p=.454, BF=.387). However, participants took significantly longer to respond “Yes” in Experiment 4 compared to Experiment 3 (β=-.255, z=-6.175, SE=.041, p<.001). This slowdown is partially consistent with a cost for relevance inferences; when “Yes” responses endorsed the inference interpretation, participants were slower to respond than when “Yes” did not endorse the inference. The lack of a difference in time to respond “No” is puzzling; if relevance inferences are costly to compute we would have expected to see greater response times to “No” when that is the inference-endorsing response rather than the literal response.

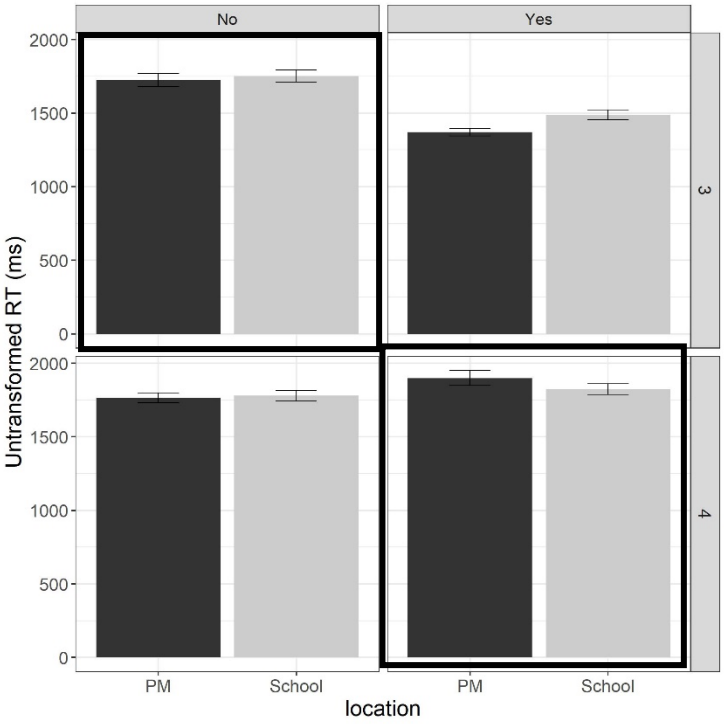

Figure 11. Response times across Experiments 3 and 4. Highlighted panels correspond to inference-endorsing responses. Experiments presented horizontally with response type presented vertically.

Pooled reading times

At the suggestion of a reviewer, to further explore the immediacy of computing relevance inferences we pooled the reading time data across both studies and reran the reading time analysis based on the response they provided. As shown in Fig. 12, the pooled data shows the pattern present numerically in Experiment 3 and significantly in Experiment 4, whereby reading times were slowest when participants went on to endorse the inference in the knowledgeable speaker condition (School).

However, the analysis shows only a main effect of Location (longer reading times for School; β=.021, t=2.097, SE=.010, p=.036) and no significant effect of Response (β=-.012, t=-.997, SE=.012, p=.319, BF=0.077) nor a significant interaction (β=.034, t=1.597, SE=.021, p=.110, BF = .015). Thus, despite the numerical pattern, the data is not reliably showing that participants are drawing an inference before the prompt question.

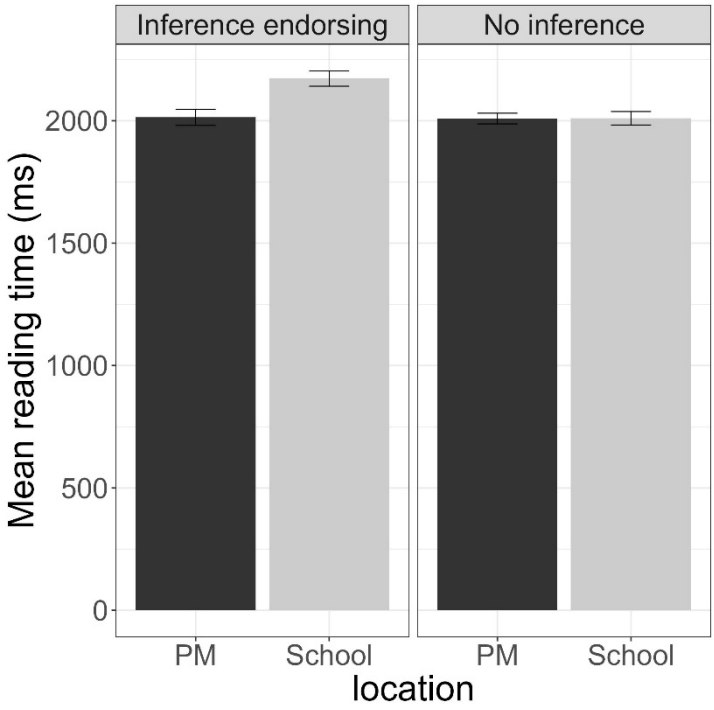

Figure 12. Pooled untransformed sentence reading times by Location and response given.

General discussion

A challenge faced by language users is recognising a speaker's intentions in producing particular utterances. A comprehender can either interpret a speaker's utterance transparently as conveying no more than the explicit signal produced or a comprehender may go beyond the transparent mapping and infer additional meaning. The search for additional meaning is typically studied in relation to *how much* information is transmitted in an utterance and often using quite contrived, unnatural contexts. In the present work we approach the question of inferencing by examining how more naturalistic utterances may prompt a comprehender to search for additional meaning and draw an inference.

Across four experiments, we asked how linguistic and extralinguistic speaker-specific characteristics contribute to the processing of utterances and the potential inferences that arise.

The first two experiments showed that when confronted with trivial utterances, comprehenders will often derive additional meaning in order to reconcile their expectations for cooperative communicative contributions. Participants look for a reason as to why a comprehender produced that utterance and draw relevance inferences (Grice, 1975; Kravtchenko & Demberg, 2022; Levinson, 2000) and this occurs without explicit cueing. Thus, a speaker's choice to speak is met with an expectation from the listener that the speaker is going to provide a suitable contribution (Rohde et al., 2021; 2022; Sperber & Wilson, 2004). When this expectation is not met, however, comprehenders work to recover a possible goal or reason that accounts for the speaker's production of a trivial utterance, akin to reasoning about why a speaker produced an underinformative utterance. As seen in other classes of inferencing, relevance inferences occurred most frequently when produced by a cooperative and knowledgeable speaker.

We then turned to the processing of relevance inferences. We measured reading time along with response times and the data is inconclusive. The reading time data suggests that

computing a relevance inference may not require additional processing during initial sentence comprehension. Numerically there is a slowdown in reading time when participants go on to endorse the inference, however this is not a statistically robust finding. Though, we do see a cost in the response times later during the explicit inferencing task when participants are prompted to consider the inference.

Taken together these findings suggest (1) that the inference may not be computed while processing the sentence itself and (2) the computation of such an inference is not cost free.

Speaker specificity

A robust effect of Speaker Knowledge was seen across all experiments: Inferencing was more frequent when the speaker was knowledgeable about the location they were talking about (when Suzy was talking about the School rather than the Prime Minister's office). In particular, participants inferred that what was asserted as the current situation did not generally hold: When told "the library walls are blue" participants inferred that the walls *used to be different* (Bergen & Grodner, 2012; Moty & Rhodes, 2021; Orena & White, 2015; Yoon, Jin, Brown-Schmidt, & Fisher, 2021). A speaker being considered knowledgeable of the topic they are speaking about increases the expectation that they will produce relevant content about that topic. If there were no effects of Speaker Knowledge, we would have expected to see similar response rates across both locations.

Interestingly but perhaps not surprisingly, when such utterances were prefaced with an explicit linguistic cue, "Hey, guess what," we did not find any increase in inferencing. This finding indicates that the decision to utter anything at all is a strong enough cue that a speaker intends to communicate cooperatively (Bergey & Yurovsky, 2022; Bohn, Tessler, & Goodman, 2019; Brown & Dell, 1987; Grice, 1975; Kreiss & Degen, 2020; Sperber & Wilson, 1995/2004).

Furthermore, speaker style influenced the rate at which participants derived relevance inferences. Relevance inferences were endorsed more often when the speaker was characterised as being "quiet" and "not very chatty". The decision to speak may carry more weight from a reticent speaker than a speaker who is talkative, yielding an expectation that a quiet speaker will only speak if their contribution is sufficiently important for the current situation (Carston, 2004; Grice, 1975; Levinson, 2000; Sperber & Wilson, 1995). This characterisation is partially supported by our findings; there were greater rates of inferencing when the speaker was described as quiet.

Surprisingly, however, this finding was rooted in a difference in rates of inferencing about the unfamiliar location. Participants were more likely to derive a relevance inference for the unfamiliar location with the quiet speaker when compared to the normal speaker. For the unfamiliar location, any utterance could be an acceptable contribution since the speaker is unlikely to have the knowledge of what that location was like in the past. Thus, utterances about the unfamiliar location should be interpreted as transparently conveying a fact about the location. In checking the speaker's familiarity with the locations, participants agreed that Suzy was less familiar with the Prime Minister's office than with her School and thus should presumably not have knowledge about the previous state of this location, i.e. she would have no information regarding the possibility that there had been a change in that location. This unfamiliarity therefore makes inference responses in this condition infelicitous given the framing of the prompt question; inferences that the situation has changed are not licensed when the speaker has no prior knowledge of that situation. Nonetheless, we see participants

responding with the inference-endorsing response in this condition. Our post-hoc explanation for this relates to the design itself.

What inferences are being drawn?

In Experiments 1 and 2, participants were limited in the responses they could give. When asked "what do you think it was like a few months ago?" the options were to answer "same" (thereby indicating no change) or "different" (indicating there had been a change and thus that an inference had been drawn). Consequently, in the unfamiliar condition, if participants wanted to communicate that they believed there to be a reason behind the speaker's production of the trivial utterance, whatever reason that may have been, there was little they could do to indicate this. The only response available for these participants to indicate that they believed the speaker was not merely communicating transparently was "different", even if this response failed to capture the type of inference they were drawing. Participants who responded "different" may not have been indicating that they thought the situation was different before, but they wanted to convey that there was a cooperative reason for the speaker to have produced a trivial utterance. This raises an important point regarding the study itself. While the task has been framed in such a way that the possible inference participants may draw has been scaffolded by the question (inferring change), we are not suggesting that this is the exact inference that every participant made.

In the case of the present work, the relevance inferences that we discuss correspond to the decision to search for additional meaning, rather than any particular inference. In quantity inferences (such as scalar implicatures) the use of a weaker scalar term prompts the search for additional meaning. It is the lack of a specific amount of information, by using a weaker term such as "some" rather than "all", that prompts a comprehender to reconcile this mismatch in such a way that often leads to the inference of "some but not all." The utterances in the current experiment, however, do not invite a search for additional meaning in the way that an utterance containing "some" might, as there is no scale in which to search for a stronger, more informative term which the speaker chose not to use. There are no clear alternatives that the speaker could have said (beyond saying nothing at all). Regardless, we find that participants systematically endorse an inference response for trivial utterances. This finding speaks to comprehenders' desire for speaker cooperativity; even when there is no need to search for additional meaning, as these utterances could all be interpreted as simply conveying facts, participants still searched for additional meaning. This is especially surprising for the unfamiliar, Prime Minister condition. While we cannot be sure specifically what inference participants were making, the results suggest that participants were nonetheless searching for additional meaning.

Processing triviality-driven inferences

We then turned our attention to the question of when relevance inferences arise during processing; are they immediately computed or do they arise downstream? We entertained three possibilities; (1) that relevance inferences arise early during the processing of a trivial utterance, which would be reflected in an immediate processing cost being observed during reading times; (2) relevance inferences do not arise unprompted thus deferring any processing cost to a later point, i.e. when prompted; or (3) relevance inferences are not costly to compute at any point in processing.

The findings from the reading time data are inconclusive. Across both experiments we see a numerical difference in reading times whereby participants who went on to endorse the

inference interpretation in the familiar condition showed longer reading times than participants who did not (see Figs 5, 8 & 12), suggesting that computing a relevance inference occurs at the point in which the utterance is processed. However, this difference was not borne out statistically.

Based on the utterance reading time data, there are two possibilities. One is to conclude that relevance inferences are not costly to compute because we did not find a significant difference in reading times as a function of speaker knowledge in either experiment. An alternative explanation is that relevance inferences are not usually computed automatically and rather arise only when a comprehender is prompted to reason about a speaker's goals. We refrain from making any strong conclusions in either direction as the data are not able to determine between these possibilities. It is instead likely that reading times as measured in these studies were not fine-grained enough to detect any subtle differences in processing. Since participants were not given any instructions about the speed with which they were to progress through the experiment, the reading times reflect the total time spent on the whole sentence before continuing to the prompt question. Furthermore, the data was collected via the internet which may have added additional noise, preventing us from observing any patterns. A more sensitive measure, either splitting the sentence into finer-grained chunks, or using eye-tracking, for example, may be able to provide a more detailed analysis of how trivial utterances are processed and any subsequent inferences that are computed.

The response time data in Experiments 3 and 4, however, provides an indication that there is a cost associated with computing these inferences. In Experiment 3, there was a main effect of response type whereby inference-endorsing responses took longer than no-inference responses. It may be that both responses in Experiment 3 are costly for their own reasons: "No" for difficulty with negative responses and "Yes" for the cost of drawing an inference. We proposed two explanations for this finding: 1) that participants are computing triviality-driven inferences, and these are costly; or 2) that participants are not computing triviality-driven inferences and the longer response times are a consequence of responding "No". Experiment 4 addressed this confound by switching the inference-endorsing response from "No" in Experiment 3 to "Yes" in Experiment 4. Here, the main effect of response type disappears. In Experiment 4, we did not find a significant difference in response times when responding "Yes" or "No"; inference-endorsing responses and no-inference responses take an equivalent amount time.

When considering the time taken to respond "Yes" across Experiments 3 and 4, we see a difference across experiments. In Experiment 3, participants are faster to respond "Yes" than "No", but in Experiment 4, we did not find a significant difference in the time taken to respond "Yes" or "No". The crucial difference is that in Experiment 3 "Yes" corresponds to no inference being made whereas in Experiment 4 "Yes" correspond to endorsing the inference. In other words, participants are much slower to respond "Yes" when this is consistent with an inference response, therefore suggesting that triviality-driven inferences are costly to derive. Surprisingly, we did not find a significant difference in time to respond to "No" across the experiments. "No" responses take the same amount of time when it corresponds to an inference-endorsing response and when it corresponds to a no-inference response. We hypothesise that this is due to different processes occurring in each Experiment.

In the case of an inference-endorsing response, such a response is costly due to the computation of the inference (as mirrored in the increased response times to "Yes" responses in Experiment 4). In the case of a no-inference response, however, we see that when participants respond "Yes" in Experiment 3, they are much quicker to do so than when they respond "No" in Experiment 4. Consider the examples below:

(10) Was it the same a few months ago?
Yes, it was the same (no inference, Exp 3)

(11) Was it different a few months ago?
No, it was the same (no inference, Exp 4)

(12) Was it the same a few months ago?
No, it was different (inference-endorsing, Exp 3)

In (10) and (11) the responses both correspond to no inference having been drawn. Despite this, we see an increase in response times in (11) compared to (10). We argue that, while both correspond to the same response type, the process to get to this "no inference" response is markedly different. Namely, in (10) participants are able to accept the situation they were presented with, whereas in (11) participants are required to consider an alternative situation and then reject that this situation was the case. In (12), however, this process of considering an alternative situation leads to accepting that alternative which corresponds to the inference response. Thus, across Experiments 3 and 4, similar processes are involved when participants provide a "No" response, which may explain the similarities in "No" response times across the studies. On the surface this does not appear to be a particularly interesting finding; however, when considered in conjunction with the differences in response times for "Yes", whether inference-endorsing or not, this is indicative of that the computation of the triviality-driven inferences is costly.

Limitations and future directions

Despite participants' consistent inference endorsements across the four studies, it must be acknowledged that they show overall low rates of inferencing. The rates of inferencing observed are at or below 50%. While this may indicate that more often than not participants are not computing triviality-driven inferences, and they deem the utterances as sufficiently relevant, there are a number of points to suggest that this is not the whole story. Firstly, we see a systematic difference in inferencing rate across Speaker Knowledge conditions. There are significantly more inferences drawn when the speaker is knowledgeable about the location; we consistently see higher rates of inferencing in the School condition compared to the PM condition. This would indicate that, although not ubiquitous, relevance inferences are more available when the context supports the search for additional meaning. This is consistent with other types of pragmatic inferencing within the literature.

Secondly, and perhaps more plausibly, there was very minimal context provided for the whole conversation. There is a brief introduction that Suzy is "telling her dad about her day" but there is very little to indicate the goals of Suzy beyond information-giving. Participants are not told what her dad already knows, and there are no prompt questions that may give insight into the exchange. Thus, this severely limited context may decrease the likelihood of drawing triviality-driven inferences (Kravtchenko & Demberg, 2022). Indeed, if these inferences are costly from a processing perspective, in the current study there is little in the way to encourage participants

to compute triviality-driven inferences and the consequences of not drawing a triviality-driven inference is minimal. In a more explicitly communicative task, or in an interactive setting we may very well find greater rates of triviality-driven inferences when there are greater pressures for accurately retrieving speaker meaning.

When considering the items themselves, it is likely that there is a ceiling for inference-endorsing responses. A number of the items refer to changes that are infrequent and may require a considerable undertaking (e.g. painting walls or recarpeting). Thus, participants may be also using their real-world knowledge to consider the feasibility of such changes, and this may have influenced their reasoning beyond considerations of acceptable conversational contributions. In post-study feedback (see supplementary materials) a number of participants did mention considering how likely it was for particular items to have changed when responding - one participant notes that it is much easier to add flowerpots outside than to paint an office.

Finally, the effect of Speaker Knowledge has been replicated on 4 separate occasions with 4 different groups of participants (totalling around 800 participants) and with different questions; the first two studies asked participants "what was it like before" Experiment 3 asked if the situation was "the same a few months ago", and Experiment 4 asked if the situation was "different a few months ago". Despite these differences, inference rates are nearly identical across the four studies. Thus, it is unlikely that this finding is a fluke.

In the present study we purposely left the question under discussion unspecified and only told participants that "Suzy is telling her dad about her day". The minimal context did not allow for participants to build strong expectations as to what a suitable contribution would look like. Interpreting the utterances as transparently conveying facts would not be surprising, especially considering that Suzy was judged as being more familiar or equally familiar with the locations than her dad. Given Suzy's "expert" status, it is plausible to interpret her utterances as intending to convey facts about those locations rather than imparting additional information beyond the observations she utters. Furthermore, participants were not the addressees of the utterances; rather they were only "overhearers" of the exchange and as such may be less compelled to derive any inferences at all. Given the lack of context, it is even more notable to have found such a robust effect of Speaker Knowledge on rates of inferencing.

Conclusion

The experiments presented here tap into the types of reasoning undertaken by comprehenders during everyday communication and hence help address broader questions about how and when inferences are computed. Comprehensive models of pragmatic reasoning need to specify the particular factors involved in computing additional meaning and future work should aim to investigate the interactions between these factors beyond the traditional scope of scalar implicatures. The work presented here provides a starting point for furthering our understanding of listeners' inferencing. Our results extend previous work on inferencing, which typically targets specific classes of words that give rise to inferences and demonstrates that broader, systematic inferencing that can arise when addressees reason about speaker goals even in the absence of cues to pragmatic enrichment.

Acknowledgements

Work reported in this paper was supported by a Leverhulme Early Career Fellowship awarded to AR and a Moray Endowment fund grant awarded to AR & HR.

References

Arnold, J. E., Hudson Kam, C., & Tanenhaus, M. (2007). If you say thee uh- you're describing something hard: The on-line attribution of disfluency during reference comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 914e930.

Arnold, J. E., Tanenhaus, M. K., Altmann, R., & Fagnano, M. (2004). The old and thee, uh, new. Psychological Science, 578e582.

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

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H. … & Grothendieck, G. (2011). Package 'lme4'. *Linear mixed-effects models using S4 classes. R package version*, *1*(6).

Bates, E. (1974). Language and context: studies in the acquisition of pragmatics. Unpublished doctoral dissertation. University of Chicago.

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1450.

Bergey, C. & Yurovsky, D. (2022, preprint). Using contrastive inferences to learn about new words and categories.

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2022). Predicting pragmatic cue integration in adults' and children's inferences about novel word meanings. *Journal of Experimental Psychology: General*, *151*(11), 2927.

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

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes*, *28*(4), 443-467.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100(3), 434-463.

Brown, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive psychology*, *19*(4), 441-472.

Brown-Schmidt, S., & Konopka, A. E. (2011). Experimental approaches to referential domains and the on-line processing of referring expressions in unscripted conversation. *Information*, *2*(2), 302-326.

Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., & Rodd, J. M. (2017). Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition. *Cognitive Psychology*, *98*, 73-101.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review, 82*(1), 45–73. https://doi.org/10.1037/h0076248

Carston, R. (2004). Relevance theory and the saying/implicating distinction. *The handbook of pragmatics*, 633-656.

Crystal, D. (1991) *A dictionary of linguistics and phonetics*. (3rd edn.) Oxford : Blackwell.

Davies, C. & Arnold, J. (2019). Reference and informativeness: How context shapes referential choice. In: Cummins, C. & Katsos, N. (eds.) The Oxford Handbook of Experimental Semantics and Pragmatics. Oxford University Press , Oxford , pp. 474-493. ISBN 9780198791768

Davies, C., Porretta, V., Koleva, K., & Klepousniotou, E. (2022). Speaker-Specific Cues Influence Semantic Disambiguation. *Journal of Psycholinguistic Research*, *51*(5), 933-955.

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

Dubois, B. L. (1989). Pseudoquotation in current English communication: "Hey, she didn't really say it". *Language in society*, *18*(3), 343-359.

Garrod, Simon, and Martin J. Pickering. "Joint action, interactive alignment, and dialog." Topics in Cognitive Science 1.2 (2009): 292-304.

Geurts, B. (2010). *Quantity implicatures*. Cambridge University Press.

Grice, H.P (1975). Logic and Conversation in Cole & Morgan (Eds) *Syntax and Semantics,* Vol. 9, Academic Press, New York (1975), pp. 411-458

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

Grodner, D., & Sedivy, J. C. (2011). The Effect of Speaker-Specific Information on Pragmatic Inferences. In *The processing and acquisition of reference* (Vol. 2327, pp. 239-272). MIT Press.

Hirschberg, J.(1985). *A Theory of Scalar Implicature.* Ph.D. dissertation, University of Pennsylvania

Horn, L. (1972) On the Semantic Properties of the Logical Operators in English. Indiana University Linguistics Club, Bloomington, IN.

Horn, L. R. (2004). Implicature, in L. R. Horn & G. Ward (ed.), *The Handbook of Pragmatics*, pp. 3–28. Oxford: Blackwell Publishing.

Horowitz, A. C., & Frank, M. C. (2016). Children's pragmatic inferences as a route for learning about the world. *Child development*, *87*(3), 807-819.

Jara-Ettinger, J., & Rubio-Fernandez, P. (2021). The social basis of referential communication: Speakers construct physical reference based on listeners' expected visual search. *Psychological Review*. Advance online publication.

Kampa, A., & Papafragou, A. (2020). Four-year-olds incorporate speaker knowledge into pragmatic inferences. *Developmental science*, 23(3), e12920.

Kravtchenko, E., & Demberg, V. (2022). "Informationally redundant utterances elicit pragmatic inferences." *Cognition* 225, 105159.

Kreiss E, & Degen J. (2020). Production expectations modulate contrastive inference. In Proceedings of the 42nd Annual Conference of the Cognitive Science Society, pp. 259–65. Austin, TX

Laver, J. (1974).Semiotic aspects of spoken communication. London : Edward Arnold.

Lenth, R. (2021). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.6.0.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic bulletin & review*, *9*(3), 550-557.

Loy, J. E., Rohde, H., & Corley, M. (2019). Real-time social reasoning: the effect of disfluency on the meaning of some. *Journal of Cultural Cognitive Science*, *3*(2), 159-173.

Malinowski, B. (1923). The problem of meaning in primitive languages. In Ogden, C. K. &Richards, I. A. (eds.),*The meaning of meaning*. London : Routledge & Kegan Paul.296-336

Morey R, Rouder J (2024). _BayesFactor: Computation of Bayes Factors for Common Designs_. R package version 0.9.12-4.7, <https://CRAN.R-project.org/package=BayesFactor>.

Moty, K., & Rhodes, M. (2021). The Unintended Consequences of the Things We Say: What Generic Statements Communicate to Children About Unmentioned Categories. *Psychological Science*, *32*(2), 189-203.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13*(4), 329-336.

Orena, A. J., & White, K. S. (2015). I forget what That's called! Children's online processing of disfluencies depends on speaker knowledge. Child Development, 86(6), 1701e1709.

Papafragou, A., Friedberg, C., & Cohen, M. L. (2018). The role of speaker knowledge in children's pragmatic inferences. Child Development, 89(5), 1642-1656.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for statistical Computing, Vienna, Austria.

Regel, S., Coulson, S., & Gunter, T. C. (2010). The communicative style of a speaker can affect language comprehension? ERP evidence from the comprehension of irony. *Brain research*, *1311*, 121-135.

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

Rohde, H., Hoek, J., Keshev, M., & Franke, M. (2022). This better be interesting: A speaker's decision to speak cues listeners to expect informative content. *Open Mind*, 6, 118-131.

Rubio-Fernandez, P. (2019). Overinformative speakers are cooperative: Revisiting the Gricean Maxim of Quantity. *Cognitive science*, *43*(11), e12797.

Ryzhova, M., Mayn, A., & Demberg, V. (2023). What inferences do people actually make upon encountering informationally redundant utterances? An individual differences study. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45, No. 45).

Searle, J. R. (1992). Conversation reconsidered. *JR Searle et al*, 137-147.

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

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Shukla, V., Long, M., Bhatia, V., & Rubio-Fernandez, P. (2022). Some sentences prime pragmatic reasoning in the verification and evaluation of comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(4), 569.

Singh, R. (2019). Context, content, and the occasional costs of implicature computation. *Frontiers in Psychology*, *10*, 2214.

Sperber, D. & Wilson, D. (1995). Relevance: Communication and Cognition. Oxford: Blackwell; Cambridge, Mass. Harvard University Press. Second edition (with postface) 1995.

Sperber D, Wilson D. 2004 Relevance Theory. In: LR Horn, G Ward, eds. The Handbook of Pragmatics. Oxford: Blackwell; 2004, 607–632.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587-10592.

Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42(4)*, 1096-1104.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44(1)*, 24-31.

van Tiel, B., Pankratz, E., & Sun, C. (2019). Scales and scalarity: Processing scalar inferences. *Journal of Memory and Language*, *105*, 93-107.

Van Tiel, B., & Schaeken, W. (2017). Processing conversational implicatures: alternatives and counterfactual reasoning. *Cognitive science*, *41*, 1119-1154.

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of semantics*, *33*(1), 137-175.

Wang, S., Sun, C., Tian, Y., & Breheny, R. (2021). Verifying negative sentences. *Journal of Psycholinguistic Research*, *50*(6), 1511-1534

Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology, 52*, 133–142. https://doi.org/10.1111/j.2044-8295.1961.tb00775.x.

Wearing, C. J. (2015). Relevance theory: pragmatics and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 87-95.

Yoon, S. O., Jin, K. S., Brown-Schmidt, S., & Fisher, C. L. (2021). What's New to You? PreSchoolers' Partner-Specific Online Processing of Disfluency. *Frontiers in psychology*, *11*, 3739.

Yoon, S. O., Koh, S., & Brown-Schmidt, S. (2012). Influence of perspective and goals on reference production in conversation. *Psychonomic Bulletin & Review*, *19*(4), 699-707.

Žegarac, V., & Clark, B. (1999). Phatic interpretations and phatic communication. *Journal of Linguistics*, *35*(2), 321-346.

Zhang, J., & Wu, Y. (2023). Epistemic reasoning in pragmatic inferencing by non-native speakers: The case of scalar implicatures. *Second Language Research*, *39*(3), 697-729.