What inferences do people actually make on encountering a redundant utterance? 
An individual differences study

Margarita Ryzhova*, Alexandra Mayn, Vera Demberg

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

Abstract

Upon hearing an utterance mentioning an activity which is highly predictable in the given context (for example, if the utterance mentions the actor eating in the context of going to a restaurant; paying the cashier in the context of going shopping; or bringing their swimsuit in the context of going swimming), subjects have been shown to rate the probability that this activity usually occurs lower than if it is not explicitly mentioned. Previous work assumed that by doing so, subjects try to make sense of the redundant utterance (as its content can be normally inferred based on the world knowledge – namely, it is well known that people usually eat when going to a restaurant) via pragmatic inferencing. Thus, subjects were assumed to come up with an alternative informative explanation of why the utterance was said (e.g., the reason to say that the actor ate in the restaurant is that usually they do not eat there and maybe usually order only some drinks). In the following study, we confirm this assumption. We also find that, just like for other pragmatic phenomena, people differ in whether they derive atypicality inferences: there are subjects who are consistently literal, consistently pragmatic, and those who are inconsistent. Their likelihood of pragmatic responding was found to be correlated with the non-verbal intelligence (IQ): individuals with higher IQ score gave more pragmatic responses.

Introduction

When reading stories about an everyday activity (e.g., about going shopping or going to a restaurant), an event that is highly associated with this activity (e.g., paying for groceries or eating) is inferred by comprehenders, even if it was not explicitly mentioned in the story (Bower et al., 1979). Uttering such an event explicitly as in (1), is hence informationally redundant (IR). Kravtchenko and Demberg (2015) showed that mentioning such events overtly triggers pragmatic inferences.

(1) Mary went to a restaurant. She ate there!

*corresponding author: mryzhova at coli.uni-saarland.de
When listeners encounter such IR utterances, they rate the probability of Mary usually eating in a restaurant lower than if it is not mentioned. This phenomenon is called atypicality inferences and it was successfully replicated in subsequent studies (Ryzhova and Demberg, 2020; Kravtchenko and Demberg, 2022). Kravtchenko and Demberg (2015) interpret such belief changes as a repair mechanism which accommodates the common ground to make the IR utterance informative with respect to the context (e.g., that Mary does not usually eat at restaurants). Consequently, the behaviour that is typically entailed by the context (going to a restaurant for eating) becomes atypical for the utterance’s referent. Atypicality inferences are highly context-sensitive and were shown to disappear in atypical contexts (e.g., when Mary is described as a person who does not like eating out) or when the target utterance refers to the event that is not highly associated with an everyday activity (e.g., “Mary went to a restaurant. She got to see their kitchen!”).

Kravtchenko and Demberg (2022) also showed that atypicality inferences are stable independently of the framing of the IR utterance. They tested three conditions of the IR utterances (with an exclamation mark as in (1), a discourse marker “Oh yeah, she ate there.”, and a plain utterance). In all three conditions of an IR utterance, the probability of Mary usually eating in a restaurant was significantly lower compared to the condition with no IR utterance, indicating an atypicality inference. The extent to which people lowered their beliefs, however, was gradual: in the absence of an exclamation mark or a discourse marker, the strength of atypicality inferences was lower. In the following study, we use the IR utterances with an exclamation mark to additionally reinforce speaker’s intentionality, thus making an atypicality inference more prominent for participants.

Following Kravtchenko and Demberg (2015, 2022), by lowering the probability of the target event, comprehenders try to make sense of redundant information by coming up with an alternative (informative) meaning for the utterance, which extends the conversational context (for example, that Mary typically goes to restaurants for drinks but does not order food). However, whether participants actually make such inferences has never been tested directly.

In the following study, we address this question by asking participants to provide an explanation for their probabilistic belief estimates.

Annotating these explanations gives us a qualitative and quantitative picture of how people accommodate redundancy and allows us to check whether the atypicality interpretations are indeed identifiable based on the typicality ratings. Compared to probability estimates, annotations yield themselves better to statistical analysis since the atypicality effect is gradual and atypicality inferences may be masked due to participants utilizing the scale differently.

We were also interested in whether participants’ biases in their interpretation are consistent and, if so, whether they are modulated by cognitive or personality traits. In order to answer the first question, we compute the proportion of pragmatic responses for each participant. To address the second question, we collect the following five cognitive and personality measures for each subject.

**Verbal working memory capacity** It has been argued that implicature derivation is effortful and therefore requires sufficient cognitive resources. For example, Yang et al. (2018) found that individuals with higher working memory capacity showed higher context sensitivity when deriving scalar implicatures. We hypothesized that atypicality inferences might to some extent draw on executive working memory resources – meaning that individuals with higher working memory capacity would derive them at a higher rate (see also De Neys and Schaeken, 2007; Fairchild and Papafragou, 2021, for a discussion).
Cognitive reflection It could be that derivation of some inferences requires overriding the literal interpretation to arrive at the pragmatic one. We found that the rate of pragmatic responding in a more formal pragmatic reference task is modulated by their performance on the Cognitive Reflection Test (Mayn and Demberg, 2022). If atypicality inferences behave similarly, individuals with higher ability to override the intuitive response would derive them at a higher rate.

Exposure to print or reading skill Scholman et al. (2020) showed that individuals with higher print exposure were more sensitive to certain context cues. We hypothesized that individuals with higher print exposure would be more sensitive to informational redundancy which would prompt them to derive atypicality inferences.

Non-verbal intelligence In order to derive an atypicality inference, participants need to reason about the interlocutor’s intention for making the redundant utterance. Also, we observed a positive effect of reasoning ability on pragmatic responding in a different pragmatic reference task (Mayn and Demberg, 2022). Our intuition is, therefore, that abstract reasoning ability may modulate pragmatic responding.

Socio-pragmatic abilities Socio-pragmatic abilities, as measured by the Autism Spectrum Quotient, have been found to correlate with pragmatic responding (Yang et al., 2018): people who are higher in autism may be less likely to put themselves in the interlocutor’s position or reason about why the interlocutor said what they did, therefore responding more literally. Our intuition is that it may also hold for atypicality inferences since those center on what the speaker intended to convey.

Methods

Participants

50 native speakers of English.

Materials

The materials were taken from Kravtchenko and Demberg (2015) and represented twenty-four stories describing different everyday situations (e.g. grocery shopping or going to a restaurant – see Table 1). Critical items contain an utterance by one of the characters stating that an activity which is highly predictable in the given context took place (with-IR condition). Control items were identical to the critical items except they did not contain the redundant utterance (without-IR condition). Filler items contained an utterance which was not informationally redundant.

Each story was accompanied by two questions. Question 1 was about the target activity typicality. Participants answered on a scale ranging from 0 (“Never”) to 100 (“Always”). In fillers, half of the questions was about the event that was not mentioned in the utterance. In Question 2, subjects had to explain the given rating.

Experimental procedure

In order to keep the original design of showing each participant only very few items (Kravtchenko and Demberg, 2015, 2022), but at the same time have more data per subject
Mary is a journalist who often goes to restaurants after her interviews. Yesterday she went to a popular Chinese place where she ran into her friend David. Later that day David ran into Sally, a mutual friend of him and Mary.

<table>
<thead>
<tr>
<th>condition</th>
<th>with-IR</th>
<th>without-IR</th>
<th>filler</th>
</tr>
</thead>
<tbody>
<tr>
<td>target activity</td>
<td>David said to Sally: “I ran into Mary leaving that Chinese place. She ate there!”</td>
<td>–</td>
<td>David said to Sally: “I ran into Mary leaving that Chinese place. She recently got a promotion!”</td>
</tr>
<tr>
<td>Question 1</td>
<td>How often do you think Mary usually eats, when going to a restaurant?</td>
<td>How often do you think Mary usually gets to see the kitchen, when going to a restaurant?</td>
<td></td>
</tr>
<tr>
<td>Question 2</td>
<td>Why did you place the slider in this particular position?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: An example of “going to a restaurant” story in with-IR, without-IR, and filler conditions.

for our analyses, the main experiment consisted of two experimental sessions (with five months in between).

A total of 24 items was used to create 8 balanced experimental lists were constructed with 3 items appearing in the target with-IR condition, 3 in the control without-IR condition, and 4 in the filler condition in each list. In the second session, the subjects saw lists of the same structure but consisting of items they had not seen in the first session.

Following Kravtchenko and Demberg (2015, 2022), we expect that subjects will assign lower ratings for question 1 in the with-IR condition, compared to the without-IR condition.

In the second session, all participants additionally completed a battery of five cognitive and personality tests (in the same order as below): Reading Span test (RSpan) (Caplan and Waters, 1999; Scholman et al., 2020), Cognitive Reflection Test (CRT) (Primi et al., 2016; Baron et al., 2015; Sirota and Juanchich, 2018; Thomson and Oppenheimer, 2016; Toplak et al., 2014), Author Recognition Test (ART) (Acheson et al., 2008; Martin-Chang and Gould, 2008), Raven’s Progressive Matrices Test (IQ) (Bilker et al., 2012), and Autism Spectrum Quotient (AQ) (Baron-Cohen et al., 2001).

Annotation procedure

Participants’ responses to question 2 from the atypicality task in the with-IR condition were assigned by two annotators to one of six categories. The annotators only saw the textual responses and not the ratings provided in question 1 so as to avoid possible bias. Subjects’ responses were classified as normal if they said that the subject probably performed the predictable activity as it is normal in the given context. This corresponds to a participant not making a pragmatic inference. If an alternative to the highly predictable activity was identified and then rejected as implausible, the tag notice_reject was assigned. atypicality corresponds to an atypicality inference. If the participant reported being unsure, their response was labeled not_sure. If multiple of the above tags could be applicable to the provided explanation, it was labeled difficult. Finally, if it
was completely unclear what the participant meant or if they misread the question, their response was labeled other. Example responses for each tag for the restaurant story are reported in Table 2.

Cohen’s $\kappa$ was $0.77$ ($p<0.0001$), 95% CI (0.7, 0.84), indicating substantial agreement. All disagreements were resolved jointly.

Results

Replication of the main effect Firstly, we show a successful replication of the pragmatic effect. As expected, the mean rating of activity typicality in the without-IR condition was higher ($mean = 86.91$, $sd = 20.25$) than in the with-IR condition ($mean = 73.77$, $sd = 30.16$). A mixed effects beta regression model of subjects’ ratings revealed a significant effect of story condition ($b = -0.69$, $z = -3.44$, $p < .001$).

Analysis of inferences Figure 1 shows that roughly 42% of instances did not show a pragmatic inference, while 51% did in fact show the expected atypicality inference (bottom panel). In the top panel, we can furthermore see that the atypicality inferences are indeed overwhelmingly correlated with a lower activity typicality estimate. Our study hence confirms that using the typicality ratings is a good proxy for measuring atypicality inferences.

Subjects’ comprehension preferences For each subject, we counted the number of annotation tags their answers fell into (see Figure 2). The histogram shows that some subjects (green subject IR labels) generally did make an atypicality inference, while others (red subject ID labels) did not usually show belief change in their answers (literal comprehenders). The group with black-colored subject IDs did not show any preferences in comprehension behavior and thus represents an inconsistent class.

Individual differences To relate subjects’ pragmatic responding to their individual traits, we calculated the number of “atypicality” annotations per participant. Next, we build an ordinal regression model of the number of “atypicality” annotations for each subject. As main predictors, we include each of the five collected individual measures.

<table>
<thead>
<tr>
<th>tag</th>
<th>inference drawn</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>no</td>
<td>Usually when you go to a restaurant, it is to eat.</td>
</tr>
<tr>
<td>notice_reject</td>
<td>no</td>
<td>After interviews Mary will be tired so she probably eats. She cannot just go to a restaurant for a drink after a long day.</td>
</tr>
<tr>
<td>atypicality</td>
<td>yes</td>
<td>Since David mentioned it, it sounds like she doesn’t always eat at restaurants. Maybe she also sometimes interviews people in restaurants.</td>
</tr>
<tr>
<td>not_sure</td>
<td>unclear</td>
<td>I’m not sure.</td>
</tr>
<tr>
<td>difficult</td>
<td>unclear</td>
<td>She goes to restaurants after her interviews.</td>
</tr>
<tr>
<td>other</td>
<td>unclear</td>
<td>He didn’t tell Sally which restaurant, he said that restaurant, as though they go there often.</td>
</tr>
</tbody>
</table>

Table 2: Examples for each tag in the annotation scheme.
Figure 1: With-IR condition. Mean typicality ratings ($\pm SE$; upper panel) and frequency of occurrence (lower panel) per annotations group.

Figure 2: Distribution of annotation tags for each subject. Subject IDs are colored with the most frequent annotation, signifying the answer preference for a subject: literal comprehender (red), pragmatic comprehender (green), inconsistent (black).
A significant effect of IQ score was found ($b = 0.7, z = 2.36, p = 0.02$), indicating that subjects with higher IQ score made more atypicality inferences.

Discussion

In this study, we were able to largely confirm the previous assumption, that a lowering in typicality ratings for the redundant activity corresponds to an atypicality inference – subjects explained their belief changes by assuming that the target event is not typical for the actor.

We also found that, just like for other pragmatic phenomena (e.g., see Heyman and Schaeken, 2015, for scalar implicatures), people differ in whether they derive atypicality inferences: there are subjects who are consistently literal, consistently pragmatic, and those who are inconsistent in their answers. Their likelihood of pragmatic responding was found to be correlated with IQ: the higher the IQ score, the more pragmatic responses subjects gave. IQ is a general measure of fluid intelligence, and its exact contribution to pragmatic responding remains to be explored further. It is possible, for instance, that deriving an atypicality inference involves reasoning about the speaker’s intention when producing a redundant utterance, and therefore may require sufficient reasoning ability.

We plan to collect data for more participants and repeat the described analyses.
References


