UNIVERSITÄT OSNABRÜCK

The interaction of modified numerals and QUDs

- A replication study -

originally by Westera and Adrian Brasoveanu (2014)

Mae Grenz
Merlin Marinova
Esma Sakalli
Johanna Venzke

Experimental Psychology Lab
Lecturer: Prof. Franke
Tutors: Ö. Özenoglu,
P. Tsvilodub, T. Anton

List of contents

0. Abstract

As a research group, we intended to replicate the study “Ignorance in context: The interaction of modified numerals and QUDs” by Westera and Brasoveanu (2014). This paper investigates context-sensitivity of ignorance references with regards to modified numerals. For this experiment, participants were asked to read a conversation between a judge and a witness and then decide whether the judge’s answer is valid or not. Due to a minimum number of participants, no significant results could be proven, hence, a recruiting of more participants is highly advised.

1. Introduction

Communication is effectively one of the most important life skills to acquire in life. Being able to deliver a message and vice versa receive information accurately for a greater understanding is essential for a successful communication between at least two speakers. On the contrary, poor communication skills may not only lead to misunderstanding but also to a misjudgment of a situation or a character.

Grice (1989) observed that an utterance in a conversation is usually stronger than its literal meaning and that this additional meaning is inferred. Essentially, inferring means to connect prior knowledge to a conversation to create meaning beyond what is directly said. According to Grice, there are two types of inferences: the scalar implicature, the inference that the hearer certainly has sufficient and correct information about what is actually meant by the literal meaning of the utterance, and the ignorance inference, the inference that the hearer does not understand the suggested, additional meaning coming along with the literal meaning. In the case of ignorance references, Grice states that this is the case when the hearer cannot determine which of the two disjuncts in an utterance of a disjunctive sentence is true.

Due to the importance in the field of linguistic research, ignorance in context has been researched by many linguists before, specifically accounting for several different theories about the implicatures of modified numerals. The reason as to why modified numerals are more looked into is because they tend to give rise to ignorance inferences more often (Cremers et al., 2021). More specifically, modified numerals might make it more difficult to conclude what was not necessarily expressed or strictly implied in an utterance but was still suggested after all. But even in this discussion, there appears to be disagreement regarding the differences between the modifier types, the importance of context-dependence and the question through which modifier type exactly or more prominently an ignorance reference is triggered (Cremers et al., 2021).

While Nouwen (2010), for instance, claims that the superlative modifier *at least* triggers an ignorance whereas the comparative modifier *more than* does not, Westera and Brasoveanu argue for a context-sensitivity in the ignorance implicatures for both modified numerals: ignorance inferences for both modifier types, namely *at least* and *more than*, depend on whether a precise answer was required to a question under discussion (QUD) or not. Westera et al. go even further by creating different types of QUDs to explore which question exactly needs one answer (QUD types HOWMANY is sufficiently fixed, type POLAR and WHAT are insufficiently fixed). Hence, the authors of this paper claim that the QUD type HOWMANY requires an exact answer, WHAT does not, which is why higher validity judgements are expected with the QUD type WHAT and lower validity judgments for the QUD type HOWMANY.

To conclude, there is no contrast between the two modifier types (MOD) COMP and SUP in sufficiently fixed QUDs since ignorance references do not depend on the modifier type when the context is sufficiently fixed. When the context, however, is underspecified or insufficiently fixed, then there will be a contrast between the two modifier types of COMP and SUP regarding the rise of ignorance references. The contrast of the latter can be explained by many linguists, among others the theory of Kadmon & Roberts (1986), that the audience is left to guess what the implication of an utterance is. Since some contexts are easier to imagine than others (Westera & Brasoveanu, 2014), different modified numerals (i.e., at least vs more than) can give room to a different interpretation of the context, possibly leading to an ignorance reference.

1.1. Hypotheses

After having examined previous acknowledged work and that of the paper "Ignorance in context: The interaction of modified numerals and QUDs" by Westera and Brasoveanu, the research group set up three following hypotheses to discuss the interaction of modified numerals and QUDs in terms of ignorance in context for a replication of the experiment of Westera et al.:

I. *Hypothesis:* Ignorance inferences for both modifier types are context-sensitive. More specifically, ignorance inferences depend on whether the question under discussion (QUD) requires an exact answer. While the QUD type HOWMANY requires an exact answer, WHAT does not. Therefore, for QUD type WHAT, we expect higher validity judgements, and lower validity judgements for QUD type HOWMANY.

II. *Hypothesis:* A contrast in ignorance between the two modifier types of comparative (COMP) and superlative (SUP) will only arise if the context (QUD) is insufficiently fixed. In sufficiently fixed QUDs, ignorance inferences do not depend on the modifier type, i.e., there is no contrast between the two modifier types of COMP and SUP.

As one can notice, there were a few changes made to the hypotheses in comparison to the preregistration report. After revising the original paper one last time, our research group noticed that in the preregistration report, we had not specifically expressed what type of QUD required an exact answer and what type did not (Hypothesis 1). Furthermore, there was no indication in the preregistered Hypothesis 1 about what kind of judgements our research group would expect for the different types of QUDs. Since our research group aimed to investigate the validity judgements to begin with, it would have been only incomplete to not state the expected results in Hypothesis 1.

Another change was made to Hypothesis 2: In the preregistered report, our research group only wrote down that there would be a contrast between the two modifier types in case of an insufficiently fixed context. However, we had not explicitly written down what would be the case for sufficiently fixed QUDs, namely that the modifier types would not differ from one another. Since the truthfulness of the first part of

Hypothesis 2 does not necessarily logically implicate the second part, it was of importance to explicitly note that there would be no contrast between the modifiers with sufficiently fixed QUDs.

To test the above-described hypotheses, our research group had participants read a judge's question, then read the witness's answer, to the question under discussion, which included a certain type of a modifier. Finally, the participants were acquainted with the judge's conclusion which was always an ignorance inference. Now they needed to decide on a 5-point Likert scale whether the judge's conclusion was justifiable or not.

However, due to time restrictions, the research group could not additionally conduct an exploratory analysis to investigate the reading times for every word that the participants took reading the witness's answer. The hypothesis would have been that the stronger the ignorance references are, the higher the reading times become.

2. Method and Materials

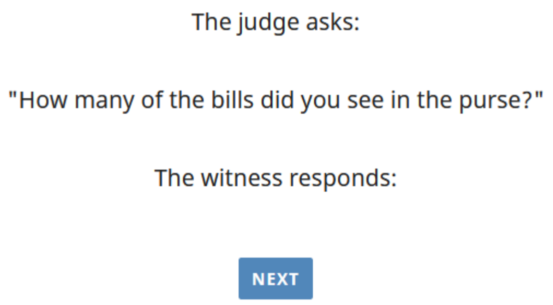
2.1. Participants

The number of participants, that we conducted for our study, was ten. None of them were excluded at the end of the experiment only for reasons of not fulfilling the requirements set by the experimenters. The requirements were following: Obviously invalid inferences must be answered with a rating of 1, otherwise the reaction will be considered as not truthful. However, in order to verify the collected data, we will only exclude data from a participant who has reacted untruthfully on more than eleven obviously invalid sentences. Reaction to obviously valid, plausible and implausible sentences will not affect data exclusion.
50% of the participants were female. All of them were in the range of 21 and 65 years old. The inclusion criteria for recruiting the participants was for them to be confident in English with a basic level of understanding, since our materials were not very hard to comprehend. The participants were reached via email to the Cognitive Science Community Server as well as through personal contacts. First, the participants read our beginning screen which stated that they could leave the experiment anytime without giving any explanation; they also had the chance of acknowledging that they were not compensated with VP hours or money for contributing to our experiment. By clicking the button "Next", the participants gave their consent for data storage. In total, they needed 30 to 45 minutes to go through all 108 stimuli, consisting of items and fillers and so to participate in our research.

2.2. Experimental Procedure

Firstly, the participants saw a welcoming screen which was followed by an instruction screen explaining the process of the experiment. Our research group conducted exactly one experiment which included participants reading a conversation taking place in a courtroom between a judge and a witness, hence, the atmosphere and item construction were made clear. The three components of an item and of a filler were always presented in this exact order on three separate screens as shown here:

First screen: Judge's question type (QUD):

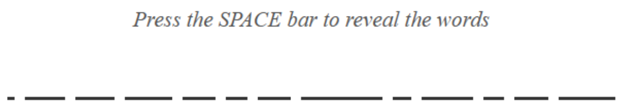


The judge asks:

"How many of the bills did you see in the purse?"

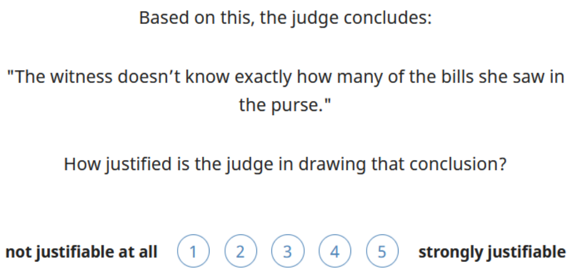
The witness responds:

Second screen: Witness' answer type (MOD):



Press the SPACE bar to reveal the words

Third screen: The ignorance inference of the judge:



Based on this, the judge concludes:

"The witness doesn't know exactly how many of the bills she saw in the purse."

How justified is the judge in drawing that conclusion?

not justifiable at all ○1 ○2 ○3 ○4 ○5 **strongly justifiable**

The second screen, containing the witness's answer, involved a self-paced reading task where participants were asked to read the answer word-by-word. Here, the SPACE bar revealed the next word while also hiding the preceding one. The reading time was recorded for each word. The third screen was used to measure how justifiable the participants think the judge's conclusion is. Participants were asked to judge the validity of the judge's conclusion on a 5-point Likert scale (1: not justified at all, 5: strongly justified). After leaving their vote on the scale, participants continued with the next item or filler.

The order of the procedure was as follows: 6 items of one condition, 12 fillers, 6 items of another condition, 12 fillers and so forth until each participant saw all of the 36 items and 72 fillers, which were in a randomized order for each participant. In total, we had 6 Lists made with a Latin square, which were rotated, so if participant 1 saw list 1 participant 2 saw list 2 and so on, which means that participant 7 saw list 1 again.

2.3. Implementation

To create the 6 lists, a double Latin square design was used. One Latin Square randomized the order of the conditions across the lists. Another Latin Square randomized the items within one list, satisfying the criteria mentioned in the Materials section. The 6 lists were hardcoded and the participant was assigned to them randomly by the sample function of lodash.

For each participant, the 72 fillers were randomized using a version of the Fisher-Yates-shuffle provided by the shuffle function of the lodash library.

Finally, the item and filler trials were mixed following a block design. 6 different items of one condition were followed by 12 filler trials, leading to a total number of 108 trials per participant.

The experiment was realized with _magpie.

2.4. Materials

Our research group created our own stimuli based on the scheme of the authors. Each item consisted of three components: a question that is asked by a judge (QUD), an answer given by the witness (including a type of scalar modifier, which was either a superlative (at least) or comparative (at most) - MOD) and a conclusion drawn by the judge based on the conversation beforehand (always an ignorance inference). In our replicated experiment, the factor QUD had three levels and MOD two levels, resulting in a total of six conditions. QUD ranged over {POLAR, WHAT, HOWMANY} (reference level: POLAR). MOD ranged over {COMP, SUP} (reference level: COMP).

Here is an example of how each of the types could look like:

- Judge's question type (QUD):

POLAR: Did you find at most ten of the coins in the wallet?
WHAT: What did you find in the wallet?
HOWMANY: How many of the coins did you find in the wallet?

- Witness' answer type (MOD):

SUP: I | found | **at** | **most** | ten | of | the | coins | in | the | wallet.
COMP: I | found | **less** | **than** | ten | of | the | coins | in | the | wallet.

- The ignorance inference of the judge:

The witness doesn't know exactly how many of the coins she found in the wallet.

While creating our stimuli, we also only used the verbs "see, find and hear" like the authors of the original study since we wanted to keep our replication study as similar as possible. The reason as to why these three verbs were used can be explained by the motivation of wanting to portray the witness as an authority over her own perception. The numeral, we chose in all our sentences, was also ten, identical to the original study, with the explanation that ten is a commonly used, round number which is not oddly linked to a specific event (like the superstition of thirteen bringing bad luck). Furthermore, the number ten is a countable number, but still a big enough number to be able to be uncertain about the precise count. Our research group also adapted the courtroom setting in order to keep the situation as natural as possible while asking very similar questions multiple times. This helped decrease the chance of participants worrying about the witness lying or even hiding information. Another factor we adapted from the original study was how all of the item parts, such as the questions, answers and conclusions, contained prepositional phrases (e.g., “ten of the coins”). This allows us to win time as well as precisibility since effects in tasks like self-paced reading (second screen) are delayed and occur a couple of seconds after the experimental manipulation point. Finally, it is important to mention that we also used only downward-entailing modifier types.

Our research group used 6 items per condition, a total of 36 items with distributions of hear (6), see (15) and find (15). Since the original study had 72 fillers (with distributed verbs hear (12), see (30), and find (30), we also had the exact same proportions. The fillers are there in order to check if the participant is paying attention. The core of the fillers was identical to the items, however, there were a few changes that we made. The changes were as following: In most fillers, in the judge’s question or witness’s answer, the words “approximately”, “probably”, “certainly” and in most fillers’ numeral modifier, the words “only” or “nearly” were added. The fillers also had an extra condition: DID (for example: Did you only see ten bread slices on the kitchen counter?). Finally, the inference of the judge in fillers was also different: It was either “The witness considers it possible that she saw {9,10,11} of the diamonds under the bed” or “The witness thinks the number of diamonds she saw under the bed is comparably high.”

2.5. Statistical Analysis

2.5.1 Variables

Manipulated variables: We manipulate the type of question posed by the judge (QUD) and the type of scalar modifier in the witness’s answer (MOD). The QUD has three levels: it ranges over {POLAR, WHAT, HOWMANY} (reference level: POLAR). MOD has two levels and ranges over {COMP, SUP} (reference level: COMP).

Measured variables: We record the reading times per word of the witness’s answer and measure how justified the participants found the judge’s conclusion. Concretely, variable RT is a metric variable capturing reaction times; variable VALIDITY_JUDGEMENT is discrete based on a 5-point Likert scale (1: not justified at all, 5: strongly justified). However, due to time restrictions, we were not able to follow up on the RT measurements and analyze them.

2.5.2 Analysis Plan

Statistical models: We will run frequentist models with variate “VALIDITY_JUDGEMENT” (whether the participant thinks the judge’s conclusion was justifiable on a 5-point Likert scale). The validity-judgement data is presented with mixed-effects ordinal probity regression models. The participants and the items are considered as random effects while the question-and-answer conditions are treated as fixed effects. An interaction model with all two-way interactions between the question conditions (POLAR, WHAT, HOWMANY) and answer conditions (SUP, COMP) and a main-effect only model will be run. All statistical modeling reported was be programmed in R. The models are estimated using the *ordinal* R package. More details are provided in the analysis script which you can find in the associated github repository.

3. Results

Figure 1 displays the mean validity judgement scores per conditions to allow for a first overview over the data. On a scale from 1 to 5, the means ranged from 3.13 for the COMP on WHAT subset to 3.15 for the SUP on HOWMANY and WHAT subset.

Mixed-effects ordinal probity regression models were used to analyze the data, since the validity judgement variable was of ordinal shape. The models incorporated intercept random effects for participants and items and included the fixed effects of the QUD and MOD conditions.
The likelihood ratio test between the interaction and main effect only model resulted in a non-significant test result (*df* = 2, $\chi^2(2) \approx 0.514$, $p \approx 0.773$).
The main effects only model evaluated for the entire dataset showed no significant results. Nevertheless, the main effects for the SUP ($\beta \approx 0.011$, *SE* ≈ 0.011, $p = 0.328$) and HOWMANY ($\beta \approx 0.009$, *SE* ≈ 0.014, $p = 0.470$) subsets were more significant than for WHAT subset ($\beta \approx -0.0002$, *SE* ≈ 0.014, $p = 0.986$).

As well as the main - effects - only model, the interaction model did not lead to any significant test result. However, the SUP on WHAT subset ($\beta \approx 0.019$, *SE* ≈ 0.028, $p = 0.496$) resulted in more significant effects than on the POLAR ($\beta \approx 0.004$, *SE* ≈ 0.019, $p = 0.838$) or HOWMANY ($\beta \approx 0.004$, *SE* ≈ 0.027, $p = 0.898$) subset.

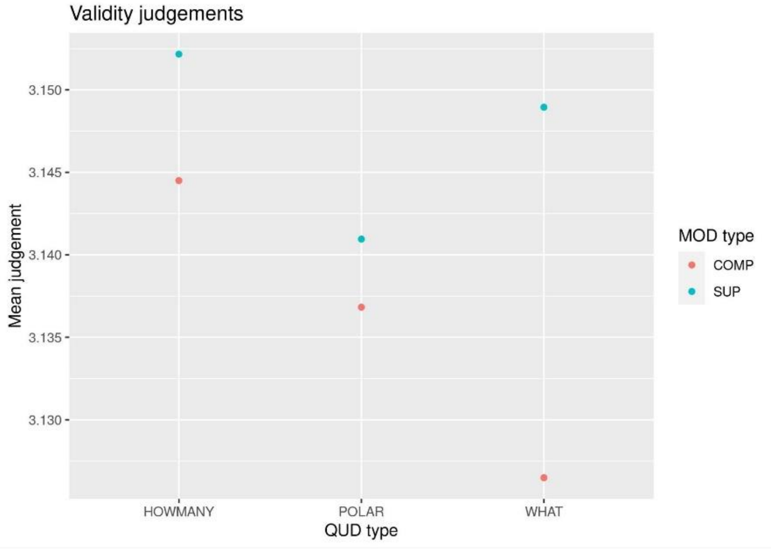


Figure 1: Mean validity judgements

4. Analyses

Since the likelihood ratio test did not result in a significant test outcome, the interaction model does not fit the data significantly better than the main-effects-only model. Evaluating the entire dataset, the differences between the ignorance inferences were not significant across the conditions. Nevertheless, the ignorance inferences in response to HOWMANY questions were slightly stronger than in response to the POLAR questions. There was no significant difference between the superlative and comparative modifier. Looking at the different subsets by question type, the contrast between SUP and COMP modifiers was again not significant. However, in response to the WHAT question condition, compared to the POLAR question condition, the contrast was somewhat strong with SUP resulting in a stronger ignorance inference than COMP.

Since none of our test results were significant, we cannot draw any conclusions regarding our hypotheses. Therefore, the inferences drawn here should be treated with precaution and only be used as indication for further research. Addressing the first hypothesis, that we expect higher validity judgements for question type WHAT and lower ones for HOWMANY, the difference between the validity judgements in the conditions WHAT and HOWMANY was in fact vanishingly small with a slightly higher judgement scores in the WHAT condition (see Fig. 1). This provides a small indication in favor of the first hypothesis but because of the small effect size, further research investigating this manner is crucial. Regarding the second hypothesis, the difference between the modifiers in response to the WHAT questions provides a first indication in favor of the hypothesis that a contrast in ignorance between SUP and COMP modifiers will only arise if an insufficiently fixed context is given. If the context is sufficiently fixed, no difference between the two modifiers is detectable.

5. Discussion

One can only presume the similarity of the outcome between this experiment and the original paper from Westera et al., because the highly significant results from the Westera et al. experiment were in this experiment only indications. Arguably, the small difference between the validity judgements for the QUD types WHAT and HOWMANY could also indicate that our data does not necessarily support the idea of context-sensitivity of ignorance references for both modifier types. Nevertheless, the second hypothesis, arguing for a contrast between modifiers in insufficiently fixed QUDs and against one in sufficiently fixed QUDs, does seem to be proven, even if only by a small difference.

The overall difference between the results from the original paper and our replicational experiment could be explained due to the very low number of only ten participants in our replicated experiment, which is underpowered and therefore possibly limits the validity of our results. The fact that the small number of participants is responsible for the non-existing significant results should be verified with further research or a replication of the study.

Hence, another replication with more participants is highly recommended. Higher participation could be achieved with a wider time frame and a small give away for encouragement for those participants that want to take part in the experiment. Since the main complaint from the participants, however, was that the experiment was quite dry and repetitive, there could additionally be more individual sentence topics for every filler question.

6. References

Cremers, A., Coppock, L., Dotlačil, J. *et al.* Ignorance implicatures of modified numerals. *Linguist and Philos* (2021).

Grice, H.P. 1989. *Studies in the Way of Words*. Harvard University Press.

Kadmon, Nirit & Craige Roberts. 1986. Prosody and scope: The role of discourse structure. In *Proceedings of the Parasession on Pragmatics and Grammatical*

Theory, Chicago Linguistics Society 22nd Regional Meeting, Chicago Linguistics Society.

Nouwen, Rick. 2010. Two kinds of modified numerals. *Semantics and Pragmatics* 3. 3:1–41.

Westera, Brasoveanu. 2014. Ignorance in context: The interaction of
modified numerals and QUDs, *Proceedings of SALT 24: 000–000*.