Replication study “Validating the paraphrase methodology” by Scontras & Goodman

Tallulah Jansen, Franziska Scharf, Louis Scheu, Stefan Warkentin

University of Osnabrück

Abstract

A problem of ambiguity in semantics and pragmatics is found in plural prediction. The present study concerns itself with the problem of distributive vs. collective ambiguity (Link, 1983, 1987, 1998; Scha, 1984; Landman, 1989a, 1989b, 1996; Lasersohn, 1988, 1990, 1995, 1998; Schwarzschild, 1994, 1996). For example, the sentence "the boxes are heavy" can be interpreted distributively, as in "the boxes each are heavy" or collectively, as in "the boxes together are heavy." Scontas & Goodman in their paper "Resolving uncertainty in plural predication" (2017) invented a methodology to assess whether the words "each" and "together" can be used to unambiguously access distributive and collective semantic interpretations and to investigate whether stubbornly distributive predicates such as "big" and "tall" consistently result in distributive interpretations. We aim to replicate their results utilizing a forced choice experiment.

Keywords: Pragmatics, ambiguity, plural prediction, collective, distributive

Replication study “Validating the paraphrase methodology” by Scontras & Goodman

An important aspect of human language is the predication (i.e. attribution) of properties to objects. When interpreting such predications, a language user is often faced with the task of resolving a number of ambiguities. In the statement “It is too hot!”, for example, the predicate “hot” can assign different properties depending on whether “it” refers to sunny weather or spicy food. Hence, to better understand the use of language for the purpose of, e.g., creating intelligent virtual assistants (Klüwer, 2011) and solving the problem of semantic role labeling (Carreras, X., & Màrquez, L. 2005), one must understand how humans resolve ambiguity with regards to predication. The interpretation of plural predications poses its own challenges. In particular, statements such as “The boxes are heavy” are subject to the so-called distributive vs. collective ambiguity of plural predication (Link, 1983, 1987, 1998; Scha, 1984; Landman, 1989a, 1989b, 1996; Lasersohn, 1988, 1990, 1995, 1998; Schwarzschild, 1994, 1996). The distributive interpretation, which holds that each box is heavy in its own right, and the collective interpretation, which holds that the boxes are heavy taken together, are both viable. Remarkably, some predicates such as “big” are not open to a collective interpretation. This phenomenon was coined “stubborn distributivity” (Schwarzschild, 2011). Although this phenomenon is well documented, it is still unclear as to why certain predicates appear to be unsuitable to be used in a collective way (Quine, 1960; Syrett, 2015; Vázquez Rojas Maldonado, 2012; Zhang, 2013).

The paper “Resolving uncertainty in plural predication” (2017) by Scontras & Goodman investigated the problem of distributive vs. collective ambiguity of plural predication and stubborn distributivity in a series of experiments. The authors proposed a methodology for disambiguating between the distributive and the collective interpretation in order to ultimately investigate which properties of predicates make humans favor one interpretation over the other (Scontas & Goodman, 2017). The present study aims to replicate their first experiment to validate their disambiguation methodology, assessing whether the words “each” and “together” can be used to exclusively force distributive or collective semantic interpretations. We also investigated whether the stubbornly distributive predicates “big” and “tall” consistently force distributive interpretations.

For this purpose, we devised an online experiment in which participants were exposed to an utterance prompt in the form of a descriptive sentence and the visual representation of two stacks of boxes simultaneously. The experimental task was to choose which stack of boxes the utterance referred to, while the referents represented either a collective or distributive interpretation of the utterance. Furthermore, exploratory analyses were conducted. We firstly investigated the effects of all manipulated variables on the time it took participants to make their choice (reaction time) and in the context of ambiguous bare utterances, we analysed the effect on referent choice for prompts including the predicate “heavy”, as we theorized to see balanced semantic interpretations. Lastly, we studied the effect of contextual cues on referent choice.

Method

Participants

We obtained a sample of N= 149. The requirements for study participation were being above the age of 18 at the time of partaking in the experiment and a strong command of the English language. English speakers were explicitly preferred. Participants were recruited via personal contacts (including public WhatsApp and Discord groups), mailing lists and the platform reddit[1]. No financial compensation was provided.

Materials

To keep in line with the original research and to set the background story of the experiment, an image from the original study (Scontras & Goodman, 2017) was used, introducing the imaginary figure Pip. He was depicted either handling a moving cart or not, depending on the scenario variant (see Figure 1). As our main visual stimulus, we used a single image throughout all experimental trials, again sourced from the original study (Scontras & Goodman, 2017). The image featured two distinct stacks of boxes, one of which was composed of five smaller boxes, while the other was composed of two larger boxes (see Figure 2). The utterance prompt we used in every trial followed the form “The boxes V1 were V2”, with V1 = “each”/ “together”/ bare (none) and V2 = “big”/ “heavy”/ “tall” in accordance with our experimental conditions.

Procedure

We applied a mixed 2x3x3 factorial design with factors 1. *scenario*, 2. *predicate* and 3. *sentence frame*. Factors had the following levels: *scenario* (“move”, “inspect”), *predicate* (“big”, “heavy, ”tall”), *sentence frame* (“each”, “together”, “bare”(none)). Importantly, condition *scenario* was measured between subjects, meaning, a subject was presented with an introductory story featuring Pip either “moving” or “inspecting” boxes at random. Conditions *sentence frame* and *predicate* were measured within subjects, presenting a sequence of nine trials consisting of random combinations of the 3x3 factor levels to the same participant.

[1] subreddits: r/academia, r/samplesize, r/takemysurvey, r/cognitivescience

The experiment was implemented using the magpie[2] architecture and hosted on netlify[3]. It exhibited the following structure:

a. Introduction
b. General instructions (varied in between subjects for condition 1)
c. Testing phase (trials varied within subjects for conditions 2 and 3)
d. Post-experiment questionnaire + Thank you

Participants were first welcomed and introduced to Pip, an imaginary figure who is a factory worker. They were informed that Pip either “inspects” or “moves” boxes through an explanatory text with a complimentary visual of him either handling a cart or without a cart, as part of the *scenario* condition. A “continue” button led participants to the subsequent page presenting general instructions and introducing the context of discourse: Pip describes boxes he had either handled or inspected to his colleague, Jim, who needs help understanding which boxes Pip refers to. A “begin” button led into the experimental trials. Pip used one of three distinct predicates, “big”, “heavy”, and “tall” for the *predicate* condition within one of three distinct sentence structures, using “each”, “together” or “bare”(none) in his utterance for the *sentence frame* condition. The utterance was displayed alongside an image of two stacks of boxes with the question posed “To which boxes is Pip referring to?” (see Figure 2). Participants each were presented with nine trials in total, offering all possible combinations of *predicate* and *sentence frame* in a random fashion. Participants were asked to choose between the referents, one of

[2] https://magpie-ea.github.io/magpie-site/
[3] https://www.netlify.com

which consistently implied a collective interpretation of the utterance (e.g. five small boxes, Figure 2, left) and one of which consistently implied a distributive interpretation (e.g. two large boxes which were together smaller than the five small boxes, Figure 2, right). They had the option to click either one of the images of the two stacks of boxes directly. After choosing one set, participants were immediately forwarded to the next trial. After completion of nine trials, participants were presented with an optional questionnaire asking “age”, “gender” (with options: male, female, other), level of education (with options: highschool graduate (diploma, Abitur or equivalent), University degree (Bachelor), Higher degree), native languages: (i.e. the language(s) spoken at home when you were a child) with an open format response section and lastly, an option to provide further comments.[4]

Our manipulated variables included *scenario,* since Pip’s access to knowledge about the boxes was manipulated between participants by having Pip either “move” or “inspect” the boxes, with or without the display of a cart respectively. We further manipulated *predicate* and *sentence frame* within subjects to again modify discourse context. For each trial, Pip utters one of the three predicates “big“, “heavy”, “tall” in combination with “each”, “together”, “bare” (none), resulting in nine randomly paired utterances in total. We measured two variables: 1. The chosen arrangement of boxes, whether a collective or distributive meaning was assigned (response/referent choice). 2. The time elapsed between stimuli onset and decision (reaction time) in milliseconds.

[4] The experiment can be found here: https://cranky-mcnulty-ca7c42.netlify.app/

Data preparation

We excluded 1 trial of a participant who completed said trial in less than 300 ms in an effort to eliminate results generated by accidental clicks or ill instruction. We further excluded 4 trials of participants who took more than one minute (60000 ms) to make their choice, for it likely did not reflect an authentic measure of their reaction time(RT). We only considered data of participants who completed the study.

The present study was concerned with the outcome of three main predictions. Our three main hypotheses were defined as follows**:**

1. All else being equal, for ambiguous sentence frames (factor-level combination: *sentence frame* = bare) featuring predicates “big” and “tall”, we expect participants to assign a distributive interpretation reflected in referent choice.
2. All else being equal, for sentence frames featuring "together" (for factor-level combination: *sentence frame* = together), in utterances such as "The boxes together were tall", we expect participants to assign a collective interpretation reflected in referent choice for the majority of trials.
3. All else being equal, for sentence frames featuring "each" (for factor-level combination: *sentence frame* = each), in utterances such as "The boxes each were tall", we expect participants to assign a distributive interpretation reflected in referent choice for the majority of trials.

If “each” and “together” act as disambiguators for plural prediction, our results should solidify the findings of Scontras & Goodman (2017) and show that distributive “each” consistently denotes a selection of the distributive referent, while a collective "together" consistently denotes a selection of the collective referent. We further expected ambiguous bare sentence frames featuring the stubbornly distributive predicates "big" and "tall" to force distributive interpretations. We expected the *scenario* manipulation to yield no effect on interpretations for just mentioned bare *sentence frames* in combination with the words "big" and "tall", since size and height are visually accessible properties.

The present study was further concerned with the outcome of exploratory predictions. For the ambiguous bare utterances, we investigate referent choice for prompts including predicate "heavy". Hypothesized by Scontas & Goodman, we adopted the expectation that *scenario* would yield an effect on "heavy", since Pip would need to physically move the boxes in one variant of the condition vs. merely inspect them. Therefore, "moves" should result in higher rates of collective interpretation. Going beyond the scope of the original paper, we investigate the influence of *scenario, predicate* and *sentence frame* on reaction time (RT).

Results

In an effort to verify the original methodology by replicating the original results, we used the same tools to analyse our data. To fit our models, we used the lme4 packages (Bates, Maechler, Bolker, & Walker, 2014) in R. Our results were coded according to whether participants chose the collective referent. Figure 3 displays the proportion of collective choices for the levels "together" and "each" of the factor *sentence frame* with a value of 1 indicating that participants chose the collective referent 100% of the time. Responses showed that an overwhelming number of participants chose the collective referent in utterances featuring "together" (72.34% in combination with predicate "big"; 73.40% in combination with predicate "heavy"; 85.21% in combination with predicate "tall") and , equally, a noteworthy number of participants chose the distributive referent in utterances featuring "each" (95.47% in combination with predicate "big"; 88.16% in combination with predicate "heavy"; 91.89% in combination with predicate "tall") (see Table 1).

To investigate the underlying effects, we fitted a mixed effects logistic regression model (Baayen, Davidson, & Bates, 2008) to predict response alias referent choice by *sentence frame* ("each","together") and its interaction with *predicate* ("big","heavy","tall") as well as trial number. Predicate "big" was coded as the reference level. The model featured random intercepts for participants (submission_id) and *scenario* (the maximal random effects structure was warranted by the data). All details of model structure and results can be found in Table 2. Model results showed a significant main effect for sentence_frame=together ($\beta = .68$, SE = .04, t = 17.03, $p < .001$), solidifying that collective "together" warrants greater numbers of collective referent choice than distributive "each". Interactions with *predicate* were not significant (predicate=heavy: $\beta = .07$, SE = 0.04, z = 1.84, p =.07; predicate=tall: $\beta = .04$, SE = .04, t = .90, p= .37). The effect of trial_number was found significant ($\beta = .01$, SE = .004, t = 2.75, $p < .01$).

In an analogous fashion, Figure 4 displays the proportion of collective choices for bare sentence frames (not featuring "together" or "each"). For the stubbornly distributive predicates "big" and "tall", we observed responses overwhelmingly in favor of the distributive interpretation ($\bar{x}$ = 92.9% for predicate "big"; $\bar{x}$ = 75.96% for predicate "tall") (see Table 3). Compared to "big", the other predicates yielded greater rates of collective referent choice in ''bare" utterances.

We modeled the underlying relations and accordingly considered responses to the bare, ambiguous utterances together with the effect of *scenario*. We fitted a second mixed effects logistic regression model predicting referent choice by *predicate* (“big”, “heavy”, “tall”), *scenario* (‘‘move,” ‘‘inspect”), and trial number. We dummy coded the *predicate* predictor, with “big” as the reference level. Our model included random intercepts and slopes for participants (grouped by trial number; the maximal random effects structure was supported by the data). The comprehensive model can be found in Table 4. We found significant main effects of *predicate* for both predicate=heavy (β = .18, SE= .04, t= 4.41, p < .001) and predicate=tall (β = .17, SE= .04, z= 4.09, p < .001).

While the effect of *scenario* was found not significant, Figure 3, however, shows a small endorsement of the expected effect on “heavy”. *scenario* variant “move” led to slightly increased collective referent choice for bare utterances with predicate “heavy” (difference of 5.66% in favor of a collective choice). Yet, *scenario* produced a small impact on all three predicates, “heavy”, “big” and “tall” in favor of a collective choice equally.

We furthermore analysed the impact of *scenario*, *sentence frame* and *predicate* on RT. Results are displayed in Figure 5. All utterance combinations with predicate “heavy” were found to have the greatest impact on reaction time. Predicate “tall” appeared to elicit shortest RTs on average, independent of *scenario*.

Discussion

Taken together, results for our first model (see Table 2) clearly support our main hypotheses that utterances featuring ‘‘together” overwhelmingly lead to collective interpretations while the ‘‘each” utterances do not, for each of the three predicates. Our results show noteworthy similarity to the original results. However, the effect of trial number, contrary to the original results, was found significant. This effect may be due to our increased sample size.

Our second model (see Table 4) indicates clear support for our main hypothesis that predicates “big” and “tall” overwhelmingly lead to distributive interpretations predicted by their stubborn distributivity with no significant influence of *scenario*. This, too, is perfectly in line with the original findings. We, equally, did not observe a significant effect of *scenario* on “heavy”, contrary to our prediction, while no effect on “big” and “tall” (since size and height are visually accessible properties) was positively predicted.

Figure 3 nevertheless shows a trend for the influence of *scenario*. We predicted *scenario* to only affect predicate “heavy” and lead to higher rates of collective interpretation, since it manipulates Pip’s interaction with the boxes and his knowledge about their physicality. The trend of higher collective choice rates for “heavy” as compared to the other two predicates was more pronounced in the original research. Our results show that the “move” scenario coupled with bare sentence frames aggregated slightly higher collective interpretations in combination with all three predications equally. Since our sample size was considerably larger, we may interpret our results as indicative of *scenario* affecting all three predicates to an equal but negligible degree. This is also supported by our second model which found *scenario* to have no significant effect.

Furthermore, we observed that out of all possible factor combinations, occurrence of predicate “heavy” had the strongest influence on prolonging reaction time (see Figure 5). Some participants provided comments stating that predicate “heavy” confused them, leaving them unable to assign either of the two meanings initially (you can find all comments in the supplementary materials referenced in **Appendix B**). Prolonged reaction time may therefore be indicative of a “startle effect”, since “heavy” used in an ambiguous utterance exhibits attributivity to both collective and distributive interpretations. We further interpret this finding as a manifestation of the effect caused by the physical invisibility of “heaviness” as compared to the visually accessible physical attributes implied by predicates “big” or “tall” (Ellis & Lederman, 1993). Future research could follow up on this finding and investigate the influence of other flexible predications on RT.

Lastly, we wish to give a conceptual outlook for future research. It has been shown that colors are suggestive of attributes (DeCamp, 1917; Payne, 1958; Payne, 1961). Since it is unclear whether the colors of the stacks of boxes confounded how they were interpreted (the color blue might appeal heavier), we think it would be of use to repeat and validate the experiment using a different color scheme. Another iteration of the experiment could use different predicates of size and shape, as they are predicted to behave stubbornly distributive (Scontas & Goodman, 2017) such as e.g. “small” and “square” in combination with a flexibly interpretable predicate such as “smooth”. As the contextual story, an imaginary figure could be depicted either looking at the objects in a storefront or moving them with a hand-held basket. Such an experimental set up could be used to investigate if there would be more contextual effect on the flexible predicate. However, researchers may have to account for the participant's prior state of knowledge about potentially depicted objects (e.g. milk cartons in Europe are packaged to contain 1 litre / ~ 1 kg of produce). Example stimuli are supplied in **Appendix C.**

Conclusion

Scontas and Goodman (2017) established a way of studying plural predication experimentally and to unambiguously access distributive vs. collective interpretations. We replicated and solidified their results showing that whenever “each” and “together” occur in an utterance, they reliably disambiguate plural predication. We were further able to replicate the finding that the predicates “big” and “tall” induce distributive interpretations as predicted by their property of “stubbornly distributive” with significant effect. Our findings therefore clearly validate the original method. Additionally, we found some evidence that flexible predicates prolong reaction time.

References

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

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. Journal of Statistical Software. http://arxiv.org/abs/ 1406.5823. ArXiv e-print.

Carreras. X., & Màrquez, L. (2005, June). Introduction to the CONLL-2005 shared task: Semantic role labeling. In Proceedings of the ninth conference on computational natural language learning (CoNLL-2005), 152-164

DeCamp, J. E. (1917). The influence of color on apparent weight: A preliminary study. Journal of Experimental Psychology, 62, 347-370.

Ellis, R.R., Lederman, S.J. The role of haptic versus visual volume cues in the size-weight illusion. *Perception & Psychophysics*53, 315–324 (1993). https://doi.org/10.3758/BF03205186

Klüwer, Tina (2011). "From chatbots to dialog systems." Conversational agents and natural language interaction: Techniques and Effective Practices. IGI Global. 1-22.

Landman, F. (1989a). Groups I. Linguistics and Philosophy, 12, 559–605.

Landman, F. (1989b). Groups II. Linguistics and Philosophy, 12, 723–744.

Landman, F. (1996). Plurality. In S. Lappin (Ed.), Handbook of contemporary semantics (pp. 425–457). Oxford: Blackwell.

Lasersohn, P. (1988). A semantics for groups and events Ph. D. thesis. The Ohio State University.

Lasersohn, P. (1990). Group action and spatio-temporal proximity. Linguistics and Philosophy, 13, 179–206.

Lasersohn, P. N. (1995). Plurality, conjunction and events. Dordrecht: Kluwer Academic Publishers.

Lasersohn, P. (1998). Generalized distributivity operators. Linguistics and Philosophy, 21, 83–93.

Link, G. (1983). The logical analysis of plurals and mass terms. In R. Bäuerle, C. Schwarze, & A. von Stechow (Eds.), Meaning, use, and interpretation of language (pp. 302–323). Berlin: de Gruyter.

Link, G. (1987). Generalized quantifiers and plurals. In P. Gärdenfors (Ed.), Generalized quantifiers (pp. 151–180). Dordrecht: D. Reidel.

Link, G. (1998). Ten years of research on plurals – where do we stand? In F. Hamm & E. Hinrichs (Eds.), Plurality and quantification. Studies in linguistics and philosophy (Vol. 69, pp. 19–54). Netherlands: Springer.

Payne, M. C. (1958). Apparent weight as a function of color. American Journal of Psychology, 71, 725-730.

Payne, M. C. (1961). Apparent weight as a function of hue. American Journal of Psychology, 74, 104-105.

Quine, W. V. O. (1960). Word and object. Cambridge, MA: MIT Press.

Scha, R. (1984). Distributive, collective and cumulative quantification. In Truth, interpretation, and information (pp. 131–158). Dordrecht: Foris.

Schwarzschild, R. (1994). Plurals, presuppositions and the sources of distributivity. Natural Language Semantics, 2(3), 201–248.

Schwarzschild, R. (1996). Pluralities. Dordrecht: Kluwer Academic Publishers.

Schwarzschild, R. (2011). Stubborn distributivity, multiparticipant nouns and the count/mass distinction. In S. Lima, K. Mullin, & B. Smith (Eds.). Proceedings of NELS (Vol. 39, pp. 661–678). Amherst, MA: Graduate Linguistics Students Association, University of Massachusetts.

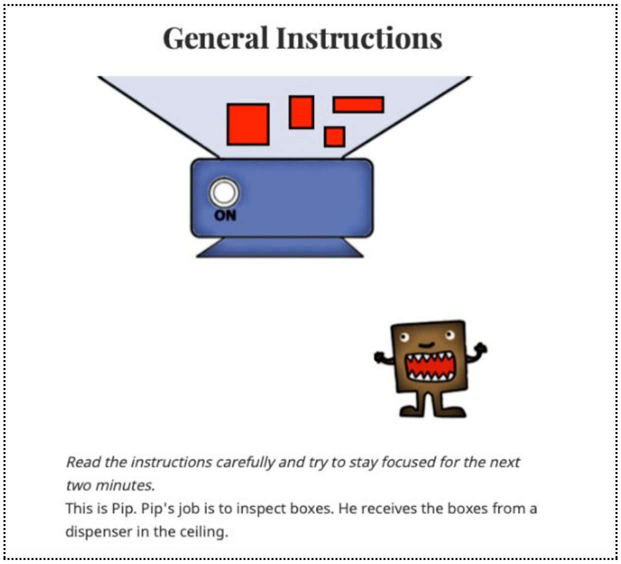
Scontras, G., & Goodman, N. D. (2017). Resolving uncertainty in plural predication. *Cognition*, *168*, 294–311. https://doi.org/10.1016/j.cognition.2017.07.002

Syrett, K. (2015). Mapping properties to individuals in language acquisition. In BUCLD 39 proceedings. Cascadilla Press.

Vázquez Rojas Maldonado, V. (2012). The syntax and semantics of Purépecha noun phrases and the mass/count distinction Ph. D. thesis. New York University.

Zhang, N. N. (2013). Classifier structures in Mandarin Chinese. Berlin: Mouton de Gruyter.

Figures

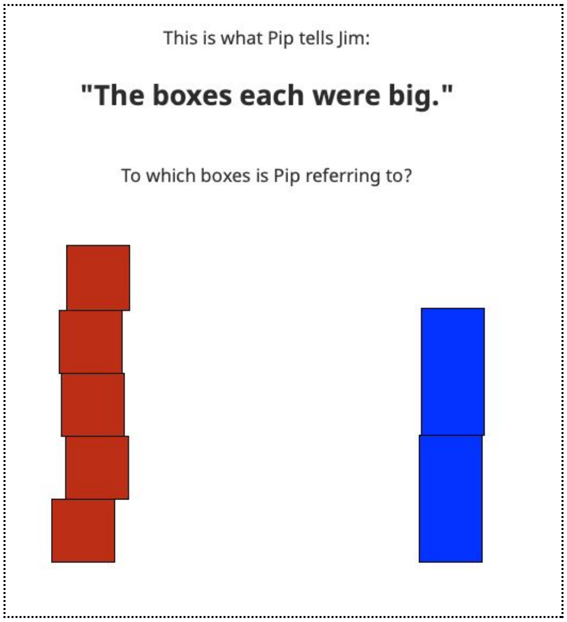


General Instructions

Read the instructions carefully and try to stay focused for the next two minutes.

This is Pip. Pip's job is to inspect boxes. He receives the boxes from a dispenser in the ceiling.

Figure 1: General instructions varying according to the *scenario* condition. This example depicts the “inspect” variant displaying Pip without a moving cart.



This is what Pip tells Jim:

"The boxes each were big."

To which boxes is Pip referring to?

Figure 2: Stimulus for one experimental trial featuring one utterance and two images resembling response options alias referent choices.

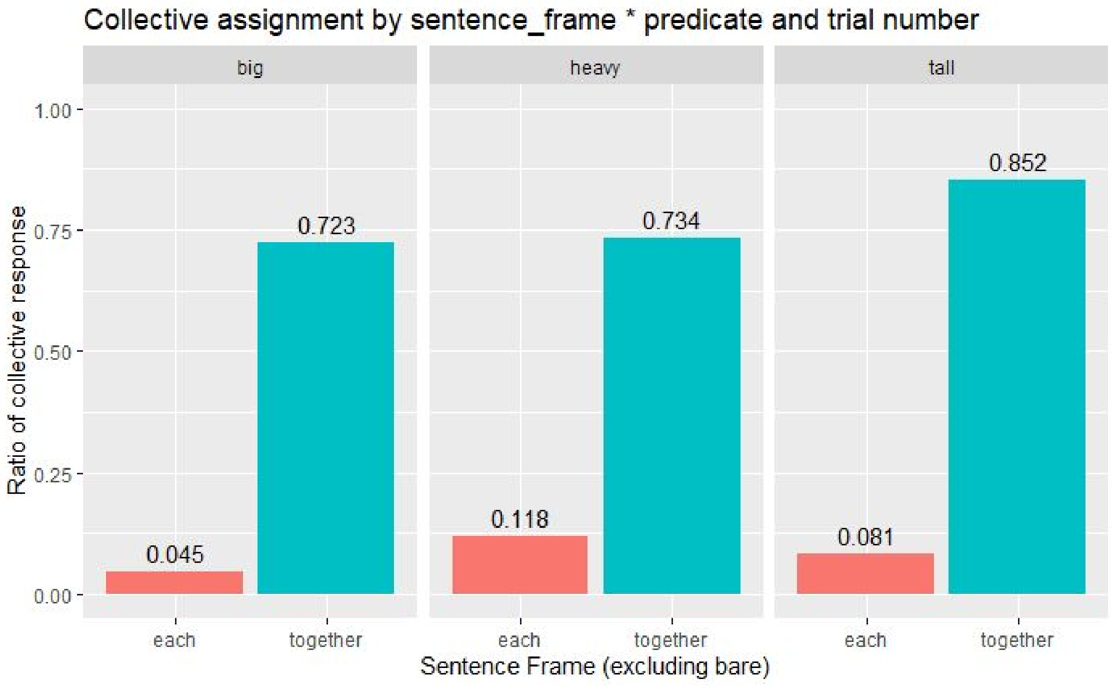


Figure 3: Proportion of collective response for *sentence frame* "together" and "each"..

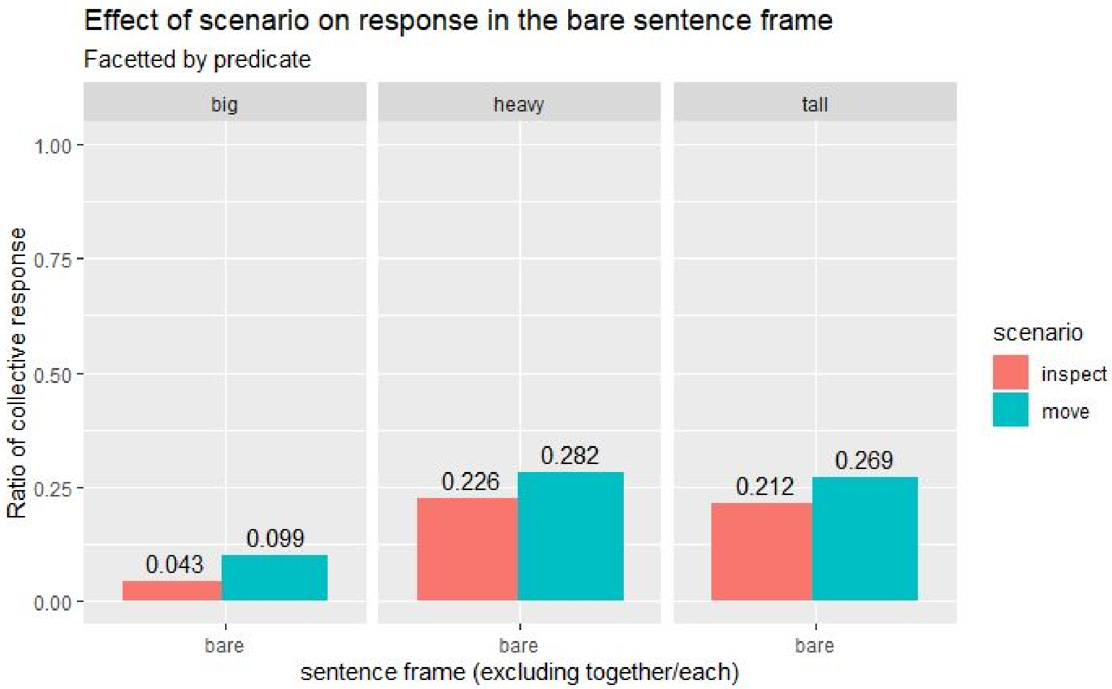


Figure 4: Proportion of collective response for *sentence frame* "bare".

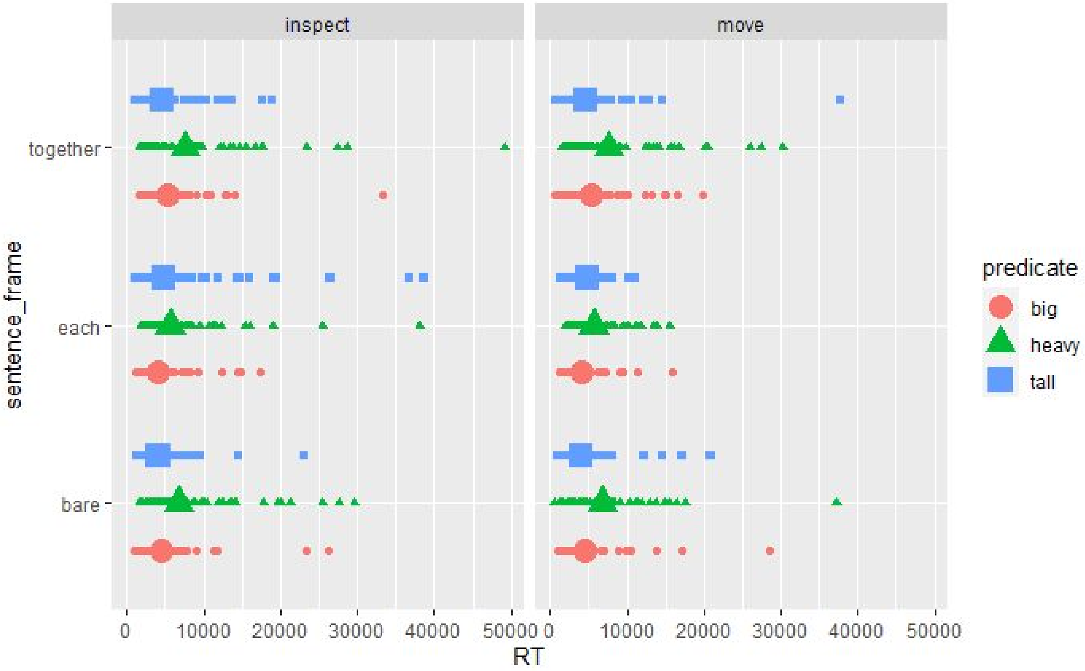


Figure 5: Reaction time by predicate and *sentence frame* facetted by *scenario*.

Tables

Table 1: Probabilities measured without influence of *scenario*.

Sentence frame	Predicate	P(collective response) (%)	P(distributive response)(%)
each	big	0.0453 (4.53%)	0.954 (95.47%)
	heavy	0.1184 (11.84%)	0.8816 (88.16%)
	tall	0.0811 (8.11%)	0.9189 (91.89%)
together	big	0.7234 (72.34%)	0.2766 (27.66%)
	heavy	0.7340 (73.40%)	0.2660 (26.60%)
	tall	0.8521 (85.21%)	0.1479 (14.79%)

Table 2: Full logistic regression model 1 from paraphrase analysis.

Total				**N=149**
Fixed effect	Estimate	Std. Error	t value	p
sentence_frame = each + predicate = big	0.033069	0.037039	0.893	0.37220
sentence_frame = together	0.678054	0.039810	17.032	< 2e-16***
predicate = heavy	0.073138	0.039724	1.841	0.06600 +
predicate = tall	0.035825	0.039739	0.902	0.36761
trial_number	0.012234	0.004448	2.750	0.00609**
Interaction sentence_frame = together + predicate = heavy	-0.062493	0.056206	-1.112	0.26656
Interaction sentence_frame = together + predicate = tall	0.092937	0.056195	1.654	0.09858

P values signified as follows : `***` = p < 0.001, ´**´ = p < 0.01, ´*´ = p < 0.05, `+`= p <0.1 (approaching significance)

Table 3: Probabilities for *sentence frame* bare.

Scenario	Predicate	P(collective response) (%)	P(distributive response) (%)
inspect	big	0.0427 (4.27%)	0.9573 (95.73%)
	heavy	0.2256 (22.56%)	0.7744 (77.44%)
	tall	0.2120 (21.20%)	0.7879 (78.79%)
move	big	0.0993 (9.93%)	0.9007 (90.07%)
	heavy	0.2822 (28.22%)	0.7178 (71.78%)
	tall	0.2687 (26.87%)	0.7313 (73.13%)

Table 3: Full logistic regression model 2 from bare utterance analysis.

Total				**N=149**
Fixed effect	Estimate	Std. Error	t value	p
scenario = inspect + predicate = big	0.0317	0.0531	0.597	0.551
scenario = move	0.0566	0.0477	1.188	0.237
predicate = heavy	0.1829	0.0415	4.409	1.45e-05***
predicate = tall	0.1693	0.0414	4.086	5.68e-05***
trial_number	0.0102	0.0078	1.413	0.160

P values signified as follows : `***` = p < 0.001, ´**´ = p < 0.01, ´*´ = p < 0.05, `+`= p <0.1 (approaching significance)

Appendices

Appendix A. Post-study questionnaire sample descriptives

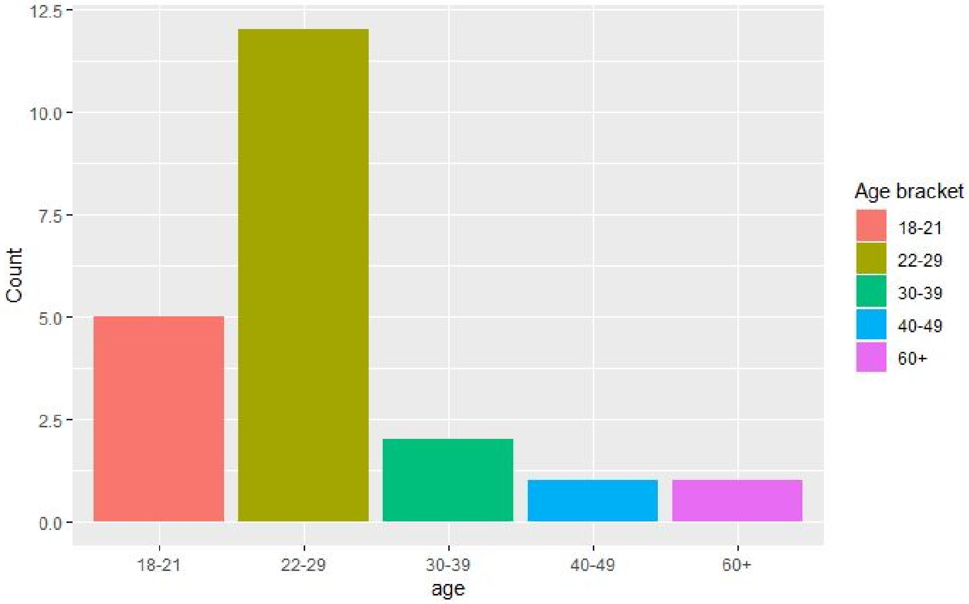


Figure 6: Results for age formatted into age brackets from the optional post-study questionnaire.

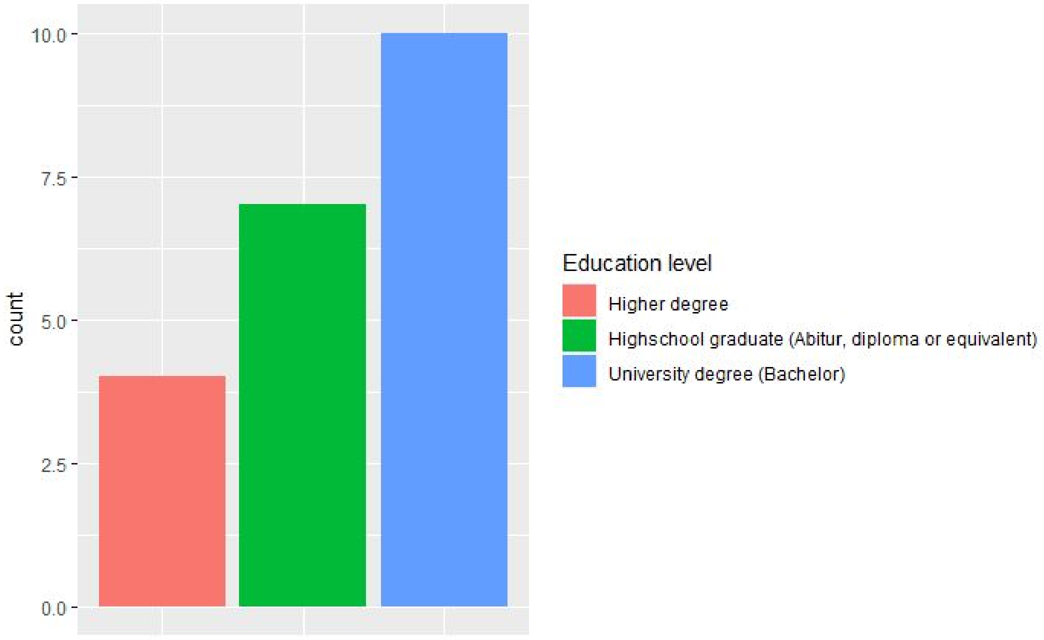


Figure 7: Results for level of education.

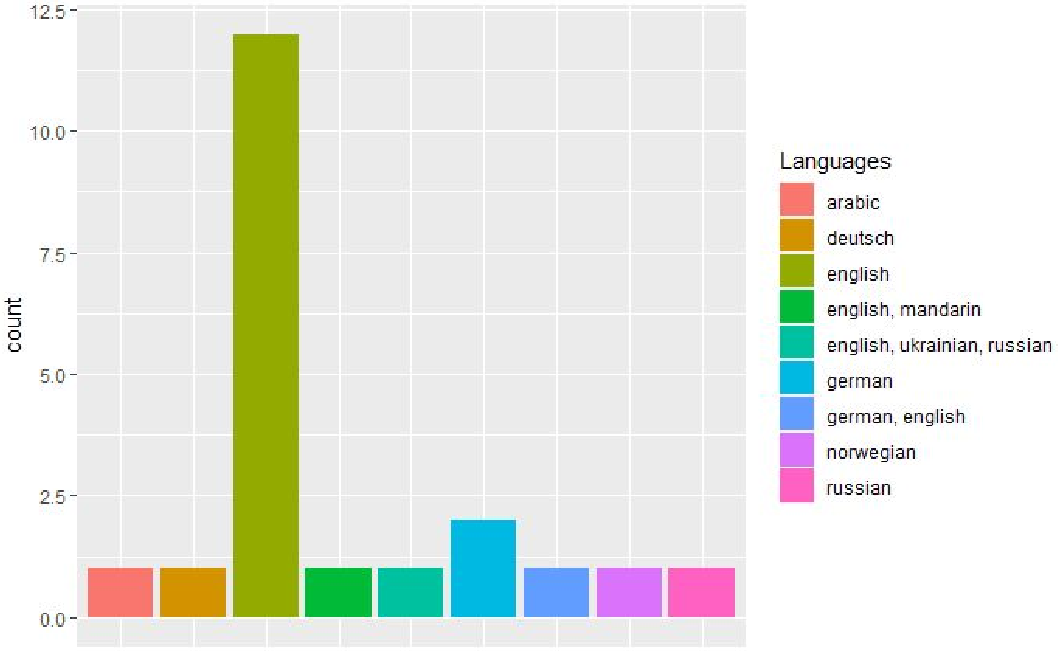
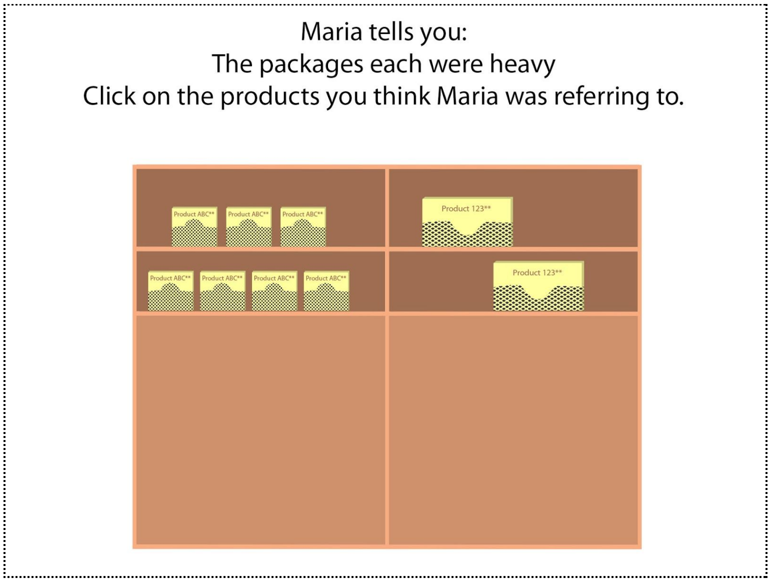


Figure 8: Results for participants’ native languages.

Appendix B. Supplementary material

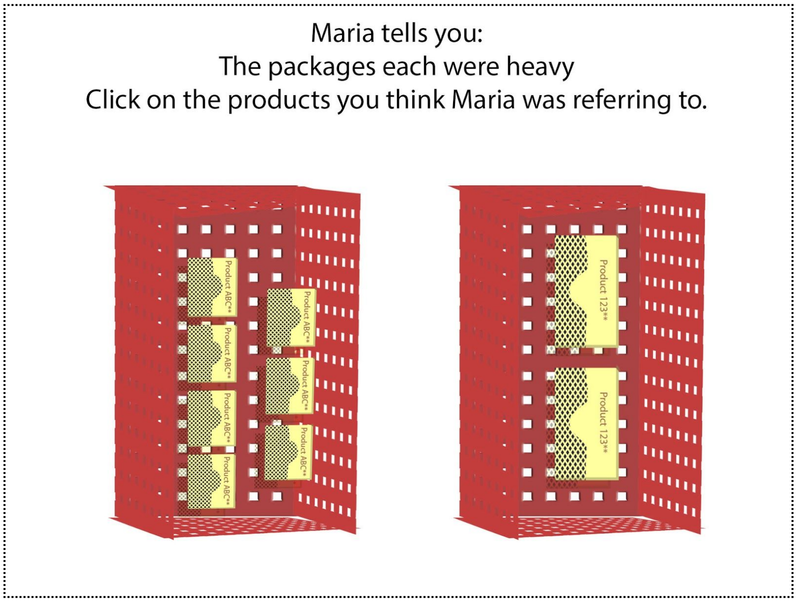
All data, materials and experimental setup can be found here: github repository.

Appendix C. Alternative Scenario Material



Maria tells you:
The packages each were heavy
Click on the products you think Maria was referring to.

Figure 9: Visual stimulus for a potential experimental design using shelves.



Maria tells you:
The packages each were heavy
Click on the products you think Maria was referring to.

Figure 10: Visual stimulus for a potential experimental design using baskets.