Communicating uncertain beliefs with conditionals: Probabilistic modeling and experimental data

Britta Grusdt (britta.grusdt@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück, Wachsbleiche 27, 49090 Osnabrück, Germany

Michael Franke (michael.franke@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück, Wachsbleiche 27, 49090 Osnabrück, Germany

Abstract

Conditionals like *If A, then C* can be used, among others, to convey important knowledge about rules, dependencies and causal relationships. Much work has been devoted to the interpretation of conditional sentences, but much less is known about when speakers choose to use a conditional over another type of utterance in communication. To fill this gap, we consider a recently proposed computational model from probabilistic pragmatics, adapted for modeling the use of conditionals in natural language, by comparing its predictions to experimental production data from a behavioral experiment. In a novel experimental approach, we manipulate relevant causal beliefs that might influence whether utterances with conditional structure are preferred over utterances without conditional structure. This is a step towards a systematic, quantitative investigation of the situations that do or do not elicit the natural use of conditionals.

Keywords: conditionals; pragmatic language use; probabilistic modeling; belief elicitation

Introduction

Formal accounts of natural language meaning are traditionally rooted in logical analysis and therefore pay attention specifically to sentential connectives. While natural language conjunctions (Blakemore & Carston, 2005), disjunctions (Simons, 2001) and negation (Horn, 1989) all feature their own respective subtleties, possibly deviating from a classical logical analysis, natural language conditionals are among the most elusive constructs to provide an account of meaning for. Given their central role in the formulation of rules and regularities, conditionals have been studied extensively in philosophy and logic (Adams, 1975; Bennett, 2003), psychology of reasoning (Wason, 1968; Evans, 1993), semantics (Stalnaker, 1968; Lewis, 1973) and pragmatics (Veltman, 1986).

The majority of existent theories on the pragmatics of conditionals concern their interpretation, a prominent one being *Mental Model Theory* (Johnson-Laird & Byrne, 2002). Other pragmatic accounts target particular phenomena observed in the communication with conditionals, such as the interpretation of '*if*' as '*only if*' (e.g. Geis and Zwicky (1971); Horn (2000)). Yet, on the production side, pragmatic accounts have remained rather vague about the reasons why speakers use a conditional sentence rather than an utterance without conditional structure. Grice (1989), for instance, argued that the utterance of a conditional commits a speaker to, what he called an *Indirectness condition*, a relation between antecedent and consequent that was yet not specified further (for recent semantic accounts, see Douven (2017); Douven, Elqayam, Singmann, and van Wijnbergen-Huitink (2018)).

A step towards a formal, predictive account for the choice of a conditional, as opposed to an utterance of a non-conditional sentence, has been advanced by Grusdt, Lassiter, and Franke (2021) who spell out a probabilistic model of utterance choice in the tradition of the Rational Speech Act (RSA) framework (Franke & Jäger, 2016; Goodman & Frank, 2016). It predicts that the speaker's utterance choice, and consequently the pragmatic listener's interpretation, depends on probabilistic beliefs about the modeled events as well as on likely causal models of the world. The model successfully captures theoretically interesting pragmatic phenomena observed in the communication with conditionals, like the listener's inferences to the speaker's uncertainty about the antecedent and the consequent or the tendency to infer '*If not A, then not C*' from the speaker's utterance of a conditional '*If A, then C*' (a phenomenon known as conditional perfection).

Here, we aim to test the RSA-model of conditionals of Grusdt et al. (2021) for its ability to predict empirical data on the speaker's choice of conditional vs. non-conditional utterances. The main challenge in applying this model to empirical data is to find a plausible setup for the set of meaning distinctions relevant for the communication with conditionals in the context of our concrete experiment, as well as a reasonable prior distribution for any assumed state space. *Prima facie*, there are at least two conceivable types of approaches: either the relevant meaning distinctions and priors that guide speaker's choices of conditionals according to the assumed model are generic and relatively independent of the concrete experimental task, or they are adapted to the specific statistics of the experimental environment. We therefore compare a model with *abstract state priors*, which assumes that certain causal relations are associated with certain kinds of probabilistic beliefs, with another version of the model, which uses *situation-specific priors* and thereby assumes participants to have acquired particular beliefs for different kinds of visual scenes shown during training.

To test these models, we collected data from a novel behavioral experiment designed to elicit human language users' descriptions of visual scenes in which utterances of conditionals might or might not be communicatively useful, depending on the relation between the represented events and their probability to occur. The experimental setup differs from most behavioral experiments investigating the meaning of conditionals (e.g. truth table tasks, acceptance ratings etc.) in that participants have the choice to actively create a variety of different conditional or non-conditional utterances. Participants describe visual scenes by creating a sentence from a set of available chunks (see Fig. 1).The visual scenes are created in such a way as to be able to systematically induce a wide range of uncertain belief states in human participants, so that, according to the model of Grusdt et al. (2021) there may be more or less of an incentive to use a conditional as a description. Concretely, we use scenes showing arrangements of objects which are more or less likely to fall, possibly as the result of another object falling (see Fig. 1 and 3), thus tapping into participants' intuitive grasp of physics to induce uncertain belief states. Other aspects of language use have been investigated by means of behavioral experiments taking advantage of peoples' intuitive understanding of physics, e.g. by Beller, Bennett, and Gerstenberg (2020), who used an RSA-model, as we do here, to investigate the pragmatics of causal language, however, not including conditionals.

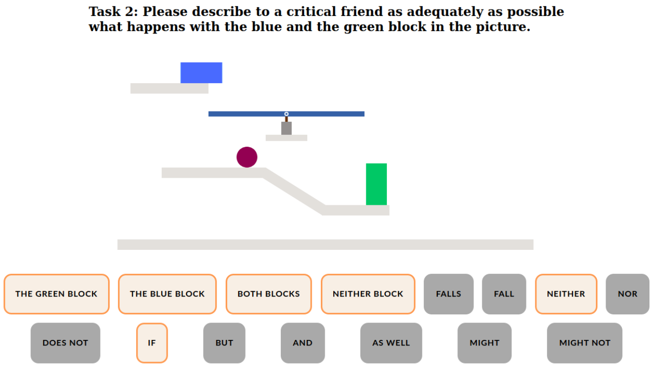

Task 2: Please describe to a critical friend as adequately as possible what happens with the blue and the green block in the picture.

Figure 1: Screenshot of a UC task trial where relation is *if2*, prior-antecedent (green block) *high* and prior-consequent (blue block) *low*. Participants chose a description of the presented scene by selecting chunks from a pre-given menu in sequence.

Experiment

Participants

We collected data from 100 English native speakers via the online crowd-sourcing platform Prolific (`www.prolific.co`). Only participants who had not participated in any of our pilot studies and had an average approval rate of at least 50% were admitted.

Materials

The experiment consisted of 15 animations in the training phase and 13 static pictures in the test phase. The stimuli from the test phase are shown in Fig. 3.

Situations differed systematically along three dimensions: 'relation', 'prior-antecedent' and 'prior-consequent', specified in Table 1. For presentation purposes, the blue block is always shown as the antecedent- and the green block as the consequent-block here, during the experiment colors were, however, randomly assigned to blocks. The prior dimensions specify how likely it is for the blocks to fall initially, without considering the respectively other block. The relation dimension specifies whether there is, intuitively, a causal relation between the two block's falling. In situations labeled as *independent*, stimuli were created such that there is likely no interaction between the two blocks, whereas in situations with *if*-labels, there most likely is. The difference between *if1*- and *if2*-trials is that in the former there is only one conceivable possibility for the consequent-block to fall, namely by the rolling ball, whereas in the latter it is *also* possible that it falls due to its position on the edge of the third block.

Table 1: Conditions for test stimuli. Letter a denotes proposition 'the antecedent-block falls', c stands for 'the consequent-block falls'. $0, L, 0.5, H$ respectively refer to an impossible event, an event that has low probability, that is expected to occur at chance level, or with high probability.

relation	prior-antecedent	prior-consequent
if1	$[L, 0.5, H]$	$P(c \mid a) = H$ $P(c \mid \neg a) = 0$
if2	$[L, 0.5, H]$	$P(c \mid a) = H$ $P(c \mid \neg a) = L$
independent	$[L, 0.5, H]$	$P(c) = P(c \mid a) =$ $P(c \mid \neg a)$, $[L, H]$

The 13 test situations were chosen to include for each relation one trial where the prior of the antecedent-block to fall is low, one where it is high and two where it is at chance level. An additional *independent*-situation was included where one block is likely to fall while the other is unlikely to fall.

The scenes shown in the test phase are slightly different instantiations of the same kind of scenes as used in the animations of the training phase. All stimuli were created with 'matter.js', a rigid body physics engine.[1] The static pictures in the test phase are 820x450 pixel screenshots of animations frozen at their initial state. The code for the experiment, all stimuli and the analysis can be found here: `https://tinyurl.com/pknmm9z9`.

Procedure

Training The purpose of the training phase was, on the one hand, to familiarize participants with the stimuli and make them acquire a good sense of the physical properties of the blocks in the simulated world. On the other hand, participants should also become familiar with the use of the sliders so that they would be able to indicate their beliefs appropriately.

In the beginning of the training phase, three comprehension questions were used to ensure that participants understood the

[1] `https://brm.io/matter-js/`

instructions, in particular the meaning of the four icons that represent the four possible outcomes of a trial (each of the two blocks falls/does not fall). In the ten subsequent preparatory trials (slider-choice trials), participants were shown pictures of slider ratings and were asked whether a given statement was an adequate description of the beliefs represented by the slider ratings.

Then the actual training phase started in which participants were shown fifteen animated situations (as described above). Before they were able to run the animations, they had to indicate how likely they believed the green and the blue block were to fall. In order to measure participant's beliefs over the two joint truth values of propositions 'the blue/green block falls', we made them adjust four sliders, one for each of the four possible outcomes (only green falls, only blue falls, both fall, neither falls). When participants had estimated the probability of all four events, their ratings were automatically adjusted to sum up to one and they were shown the result of this normalization. Participants then had the chance to update their slider ratings for as long as they liked. After each stage of selection, the current normalized probabilities were shown numerically and, as further visual help, as a blue and a green pie chart, representing the marginal probability assigned to each of the two blocks to fall. When participants were satisfied with their slider adjustments, they would click on a "RUN" button which started the animation.

Before participants could move on to the next trial, they were given feedback about which event actually occurred and how much probability they had assigned to this event. Instructions made clear that assigning a low probability to the eventual outcome might still have been a reasonable choice due to chance, and participants were encouraged to continue indicating their genuine beliefs and uncertainties. All training trials were pseudo-randomized such that the number of blocks that fall per trial was approximately evenly distributed across all trials.

Testing In the test phase, each participant saw each of the 13 test situations once in pseudo-randomized order, such that *high*/*low* and *uncertain* prior conditions were approximately evenly distributed and no subsequent trials had identical relation conditions.[2] For each situation, participants worked on two tasks in direct sequence.

The first task, called the PE (prior elicitation) task, elicited belief judgements about likely outcomes of each physical arrangement. Unlike before, during the training phase, participants were now shown static pictures instead of animations and did not get any feedback on their ratings.

The second task, called the UC (utterance choice) task, reused the same pictures shown previously in the PE task. Participants were asked to "describe to a critical friend as adequately as possible what happens with the blue and the green block in the picture". Participants were instructed not to say what they were not sufficiently convinced of (as the friend is assumed to be critical). The choice of descriptions that participants could possibly create was limited: they were shown a set of buttons with words that had to be clicked on in order to concatenate them to form sentences as shown in Fig. 1. The created utterances were shown on the screen and participants had the possibility to make corrections and optionally, they could freely type in a sentence if they did not consider any of the given possibilities to be adequate (not considered here). The experiment only allowed concatenations of sentence chunks that formed grammatical sentences.

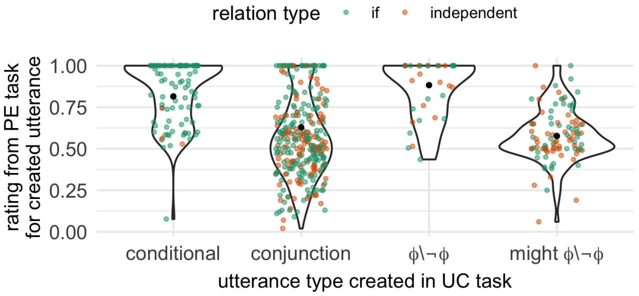


Figure 2: Densities of participants' estimations from the PE task for utterances created in the corresponding UC task in situations where the prior condition for the antecedent-block is *uncertain*.

The available utterances can be categorized into four different types. Conditionals form the category that we are mainly interested in. Further, participants could create conjunctions as they allow to explicitly mention two events with an utterance other than a conditional, and simple assertions, like 'the blue block falls' or 'the green block does not fall', describing possible outcomes for the green and blue block separately. Each of the four simple assertions could be combined with 'might' such that participants had another possibility to express uncertainty other than by using conditionals.[3] This yields a total of 20 sentences with distinct meaning, some realizable in multiple ways, e.g. 'the blue block falls and the green block falls' denote the same distinct meaning as 'both blocks fall'.

Results

Data Cleaning The entire data set of a participant was excluded if (i) they failed any of the attention check questions in the test phase (5 removed) or (ii) got more than half of the ten slider-choice trials in the training phase wrong (1 removed), and if (iii) the average squared differences between a participant's ratings in the PE task and the mean response of all other participants across all test situations was larger than 0.5 (3 removed). Three participants' data was excluded since their comments indicated that they had technical problems or difficulties with the task. We excluded six trials where partic-

[2] After each second trial, an attention check asked for the color of a block in a shown picture.

[3] Note that 'might' could not be used within conjunctions, e.g. 'blue might fall, but green does not fall' could not be created.

ipants created an utterance in the UC task that was assigned a probability of 0 in the PE task by the same person.

Behavioral Data All stimuli were shown twice in direct sequence, first in the PE task in which participants were asked to indicate their beliefs regarding the (falling) behavior of the two blocks and consequently in the UC task, in which participants were asked to create a sentence that described the visual scene. To get a sense of the relation between participants' prior ratings for a given scene and their choice of description for that same scene, Fig. 2 shows data from both tasks, namely for the six scenes where the prior condition for the antecedent-block to fall is *uncertain*. Concretely, it shows the probabilities corresponding, according to the semantics discussed below, to participants' respectively chosen utterance for scene j (grouped by utterance type), as estimated in the PE task for scene j. According to this, participants often used conjunctions or simple assertions even though they had indicated with ratings around 0.5, and even lower, to be quite uncertain about the respective outcome.[4]

To provide more detail, Fig. 3 shows the stimuli of all test situations (**A**), with the results from the PE task (**B**) and the UC task, with corresponding model predictions (**C**). As the main purpose of the experiment was to test the predictions of a computational model, further behavioral results will be discussed along with model predictions later on.

Computational Model

The computational model that we aim to test is a vanilla RSA-model (Franke & Jäger, 2016; Goodman & Frank, 2016) adapted to be applicable to communication of stochastic/causal dependencies. RSA-models are probabilistic models that formalize Gricean pragmatic reasoning: the speaker's utterance choice is predicted to depend on the utility of an utterance for communicating a state, in relation to the utility of plausible alternative utterances available to the speaker. As the relevant data we consider is for the choice of a suitable description, we focus on the speaker part of vanilla RSA:

$$P_S(u \mid s) \propto \exp(\alpha \cdot U(u;s)) \tag{1}$$

The free parameter α tweaks the extent of 'rationality' of the speaker; larger values of α correspond to stronger pragmatic inferences, i.e. the larger α, the more the speaker's predicted distribution will be peaked on the utterance with the largest utility (we set $\alpha = 3$). The utility of an utterance u for a state s, $U(u;s)$, corresponds to its degree of informativeness, defined in terms of the literal meaning of u, and is possibly attenuated by utterance costs (here set to 0). Whether an utterance u is literally true for a given state s, is defined by the denotation function $[\![u]\!](s)$ which returns 1 if u is true/assertable in s and 0 otherwise (to be specified below).

$$U(u;s) = \log P_{\text{lit}}(s \mid u) - \text{cost}(u) \tag{2}$$

$$P_{\text{lit}}(s \mid u) \propto [\![u]\!](s) \cdot P_{\text{prior}}(s) \tag{3}$$

Thus, the larger the utility of an utterance is, the easier it becomes for a literal interpreter (Eq. (3)) to distinguish the speaker's intended state from other states and so, the more likely the speaker is to choose the respective utterance as description of the given state.

It remains to specify the set of alternative utterances, their literal meaning and the definition of states. Following Grusdt et al. (2021), a state is defined as a probability table over two binary variables, namely whether or not the green, respectively the blue block, will fall in a given situation. That is, a state represents (probabilistic) beliefs about the four possible combinations of outcomes, as judged in the PE task. The shape of the probability tables is defined by a latent variable, which we will explain in more detail in the next section, in context of the the prior probability over states, $P_{\text{prior}}(s)$.

Following the experimental setup, the speaker model includes all 20 utterances which could be formed without using the free typing option in the UC task. There are four different types of utterances: conditionals, conjunctions and simple assertions, where the latter can be combined with 'might'. Utterances are defined to be literally true or assertable with respect to a given state s, when the corresponding probability derived from the probability table of s is larger than a threshold θ. The conjunction, 'both blocks fall', for instance, corresponds to the probability $P^s(B \wedge G)$ for state s. For simple assertions, with and without 'might', we consider the relevant marginal probabilities given by s, and for conditionals the relevant conditional probabilities. The literal meaning threshold θ is set to 0.7 for all utterance types except for those modalized with 'might', which are considered true when θ is larger than 0.25 (c.f. Herbstritt & Franke, 2019).[5]

Model Fitting

To derive model predictions for the data at hand, we still need to fix the state space and the state prior $P_{\text{prior}}(s)$. There are at least two *prima facie* plausible specifications. For one, **situation-specific state priors** assume that speakers will choose a description based on a (hypothetical) listener's interpretation which takes the concrete, experiment-induced statistics of the environment into account, i.e., the situations C_i that occur, and how likely each state s is for a given, concrete situation $P_{\text{prior}}(s \mid C_i)$. The empirical data from the PE task can be used to specify situation-specific state priors:

$$P^{\text{SIT}}_{\text{prior}}(s) = \frac{1}{13}\sum_{i=1}^{13} P(s \mid C_i)$$

$$P(s \mid C_i) = \text{Dirichlet}(s \mid \hat{\alpha}_i)\,,\ \text{where}$$

$$\hat{\alpha}_i = \arg\max_{\alpha} \prod_j \text{Dirichlet}(D^{\text{PE}}_{ij} \mid \alpha)$$

[4] A reason for this bias might be that participants were not able to create utterances that are stronger than assertions with 'might', but less strong than simple assertions or conjunctions (e.g. 'the blue block *probably* falls'), and so participants might have considered an assertion ϕ like 'it *will be* that ϕ' which does not require certainty.

[5] 0.7 is arguably a rather low value, the empirical average threshold is at 0.662, however, even smaller.

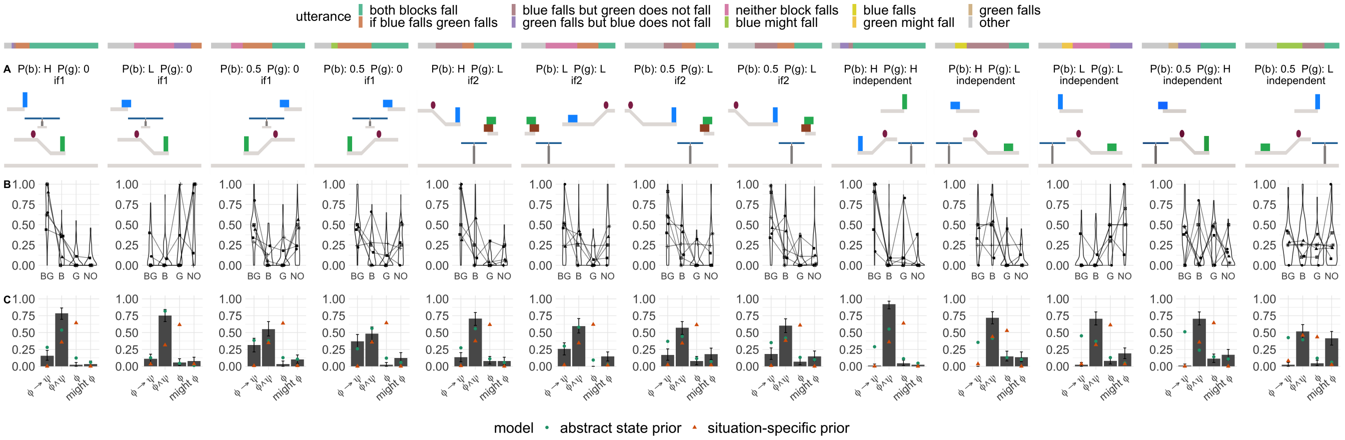
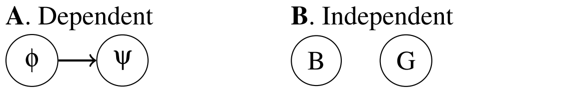


Figure 3: Results for all 13 test situations where each column shows the data from one situation. **A**. Stimuli with the relative frequency of the three utterances participants created most often in the UC task represented by colored bars above. ‘Other’ comprises all remaining utterances. **B**. Slider rating densities overlayed by observations from 6 randomly chosen participants. **C**. Proportions of utterance types produced in the UC task, plotted with the respective predictions from both models (errorbars are bootstrapped 95% confidence intervals).



$P(\psi \mid \phi) \sim \text{beta}(10,1)$ $\quad$ $P(\text{B}), P(\text{G}) \sim \text{beta}(1,1)$

$P(\psi \mid \neg\phi) \sim \text{beta}(1,10)$ $\quad$ $P(\text{B},\text{G}) \sim$

$\psi, \phi \in \{\text{B},\text{G}\}$ $\quad$ $\mathcal{N}_l^m(P(\text{B}) \cdot P(\text{G}), 5 \cdot 10^{-3})$

Figure 4: Noisy-or model for two dependent (**A**) and independent (**B**) variables, where B denotes ‘the blue block falls’ and G ‘the green block falls’. m, l refer to the upper and lower bound of $P(B \wedge G)$, constrained by the marginal probabilities, $P(B), P(G)$.

Situation-specific priors are calculated by a (theory-free) Dirichlet model. We compute the ML-estimate for a Dirichlet distribution to best predict the data from the PE task for each situation C_i.[6]

Alternatively, **abstract state priors** consider a space of weighted interpretation options for a generic case of communication about two binary variables which may or may not stand in a causal relationship. Following Grusdt et al. (2021), we consider five different causal nets, one where the two relevant binary variables are independent, and one for each of the four possible causal relations (positive/negative influence in each direction). Each causal net cn_i is associated with a stereotypical noisy-or situation (Cheng, 1997) that defines a conditional state prior $P_{\text{prior}}(s \mid cn_i)$, as specified in Fig. 4. Here, alternative causes are only implicitly taken into account insofar as the probability of the effect is larger than zero in the absence of the single cause. This parametrization is arguably a natural choice to make as several empirical and theoretical studies have shown that people tend to neglect alternative causes (e.g. Krynski and Tenenbaum (2007); Fernbach and Rehder (2013)). The probability tables associated with the independent causal net are generated by adding Gaussian noise to tables sampled for two probabilistically independent variables, to allow some deviation from the exact definition of independence. The overall abstract state priors are then:

$$P_{\text{prior}}^{\text{ABS}}(s) = \sum_{i=1}^{5} P(s \mid cn_i) \cdot P(cn_i)$$

where the marginal probability of the independent causal net is 0.5 and of any dependent causal net respectively 0.125.

The predictions of the vanilla RSA model about likely utterance choices in the UC task for a given situation C_i are obtained by averaging model predictions based on the assumptions that speakers ground their beliefs about s in the state priors (abstract or situation-specific) but also tailor them to the specific situation C_i which is to be communicated. A computationally efficient realization of this idea is to condition the state priors $P_{\text{prior}}(s)$ on the set D_i^{PE} of empirically observed belief ratings for situation i, so that:

$$\text{Prediction}(u \mid C_i) = \sum_{s \in D_i^{\text{PE}}} P_{\text{prior}}(s \mid D_i^{\text{PE}})\, P_S(u \mid s)\,.$$

Results

According to the theoretical predictions of the model from Grusdt et al. (2021), speakers’ utterance choices should depend on the prior probabilities of the modeled events and, implicitly, on the relation among them. That is, in terms of our experiment, participants’ utterance choices should be influenced by the prior and relation of the the stimuli which were

[6] The family of fitted Dirichlet distributions seems to explain the aggregate data well enough ($p \approx 0.78$ based on MC-simulations with log likelihood as test statistic), even though there are four situations where the respective Dirichlet distributions did not fit well: $p \in (0.001, 0.002, 0.005, 0.007)$, for all others $p >= 0.09$.

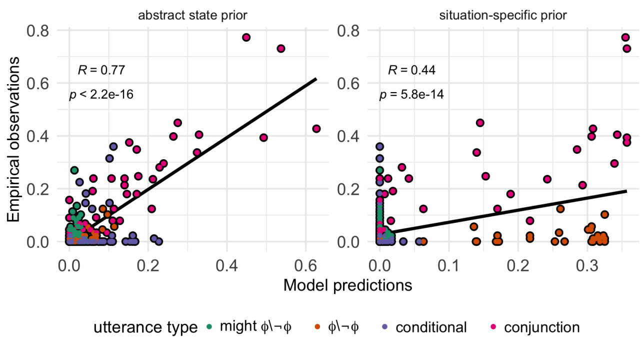


Figure 5: Overall correlation between model predictions (*left*: abstract state priors, *right*: situation-specific priors) and empirical observations, for each situation and utterance. For better readability, color code is with respect to utterance types, not single utterances.

designed to elicit different belief states. In particular, this model, i.e., the model using abstract state priors, predicts that (H1) conditionals are produced more frequently in dependent conditions *if1* or *if2* than in the *independent* conditions. It also predicts that (H2) conditionals are more likely when the prior probability of the antecedent-block is *uncertain*, because otherwise the model predicts a higher probability of a stronger utterance, i.e., a conjunction or a simple assertion.

To test these predictions, we ran a logistic regression model using the R-package brms (Bürkner, 2018) with varying intercepts and slopes per participant to predict participants' choice of conditional vs. non-conditional utterances based on the relation (independent vs. dependent) and the prior of the antecedent-block to fall (uncertain vs. confident, the latter aggregates *high* and *low*). We find strong evidence for both hypotheses formulated above, namely (re H1) a credible main effect of dependence ($P(\beta_{\text{relind}} < 0) = 1$, mean ≈ -5.6, CI=[-10.03,-2.76]) and (re H2) a credible main effect of confidence ($P(\beta_{\text{blueunc}} > 0) = 0.99$, mean ≈ 0.84, CI=[0.24, 1.38]).

In Fig. 3, the first shown stimulus of each relation-condition (columns 1, 5,9) show data of situations where both blocks have a high probability to fall. As expected, participants choose conjunctions in all three situations, conditionals are yet only observed in the dependent situations, where they are among participants' three most likely chosen utterances. The predictions from the model with abstract state priors diverge from this observed data, particularly in the independent situations, where conditionals are predicted to be chosen with a very high probability. Contrary to that, the model with situation-specific priors hardly predicts conditionals at all, but largely overestimates the use of simple assertions instead. Conjunctions are underestimated throughout by both models. One reason for this underestimation can be found considering participants' slider ratings that model predictions are based on: the four joint events are only rarely assigned values larger than the literal meaning threshold $\theta = 0.7$, yet participants tend to use conjunctions. That is, in these cases, the models fall back on predicting simple assertions, utterances with 'might' and conditionals.

Fig. 5 shows the correlation between model predictions and observed proportions for utterance choices across stimuli. On average, the abstract state prior model moderately correlates with participants' observed production behavior (R=0.77), whereas under the assumption that the speaker draws on situation-specific priors, the correlation is worse (R=0.44).

The difference between model fits of the two models suggests that the structure in the state space induced by the abstract state prior captures something relevant that the state space of the situation-specific prior model does not account for. In particular, in case of the abstract state prior, the informativeness of utterances is clearly structured: conjunctions are assertable only in few states (highly informative), simple assertions in a few more, while conditionals are assertable much more often (less informative), surpassed by utterances with 'might', which are the least informative. This structure is not induced to the same extent by the situation-specific prior.

Discussion & Conclusion

While a lot of work on the meaning and use of conditionals exists, comparatively little attention has been paid to systematic accounts of predicting whether and when human language users would actually use a conditional, rather than a non-conditional utterance, and, if they do, exactly *which* conditional sentence they prefer. This work has made a first step towards filling this gap by setting up an experiment suitable for manipulating participants' uncertain beliefs and probing their preferences for different conditional and non-conditional utterances. We tested two instantiations of an RSA speaker model to predict the empirically observed utterance choices, based on empirically measured subjective beliefs induced by various situations. Overall, it seems that with abstract state priors, as proposed by Grusdt et al. (2021), the resulting production probabilities are able to explain general utterance choice preferences.

These results are promising and pave the way for a future, more thorough investigation of theory-driven models predicting human use of conditionals. For one, the presented comparison between model predictions and empirical data indicated its ability to capture some of the observed production data, without fine-tuning its free parameters (rationality parameter α, utterance cost). Given that we observed the abstract state prior model to overestimate the use of conditionals and to underestimate the choice of conjunctions, assuming higher costs for conditionals might substantially improve model fit.

Other reasons are conceivable that might explain the model's observed overconfidence in conditionals, in particular in independent situations, where they are least expected. As these are situations where participants hardly ever produced conditionals, they seem to have grasped the intended difference in the causal relations in the different stimuli. However, it might be possible that nevertheless they produced probability tables in the PE task that are far from being probabilistically independent. If these probability tables, on which the predictions of the model are based, were more typical for other causal relationships than independence, this would justify the large predicted probability for conditionals. For probability tables that *are* stereotypical for the independent causal net, in the sense of the assumed abstract prior, the model should not predict conditionals to the observed extent.

It is also possible that the speakers' utterance choices do not only depend on their probabilistic beliefs but also take into account the underlying causal relation.[7] Here, the speaker is modeled to simply communicate her probabilistic beliefs without reference to the underlying causal relation, as the literal meaning that defines the utility of utterances, only depends on the probabilities given by a state s but not on its associated causal net. The speaker in the presented model could be extended to a speaker who does take this aspect into account when communicating her beliefs.

We chose a population-level modeling approach here. Alternatively, model predictions for datum D_{ij}^{UC} could be derived from the associated probability table that individual j gave, i.e., state $s = D_{ij}^{\text{PE}}$. However, this approach assumes that each participant will have the *exact* same mental picture of the situation each time it is presented — which may be dubious even for a presentation of the same stimulus in direct succession of different tasks. Therefore future extensions with hierarchical modeling of individual-level beliefs appear reasonable.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2340. We would like to thank Malin Spaniol and Josefine Zerbe for their support and the anonymous reviewers for constructive and helpful feedback.

References

Adams, E. W. (1975). *The logic of conditionals*. Dordrecht: Elsevier.

Beller, A., Bennett, E., & Gerstenberg, T. (2020). The language of causation. In *Proceedings of the 42nd annual conference of the cognitive science society.*

Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford University Press.

Blakemore, D., & Carston, R. (2005). The pragmatics of sentential coordination with *and*. *Lingua*, *115*(4), 569–589.

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Douven, I. (2017). How to account for the oddness of missing-link conditionals. *Synthese*, *194*, 1541–1554.

Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen-Huitink, J. (2018, may). Conditionals and inferential connections: A hypothetical inferential theory. *Cognitive Psychology*, *101*, 50–81.

Evans, J. S. T. (1993). The mental model theory of conditional reasoning: critical appraisal and revision. *Cognition*, *48*(1), 1–20.

Fernbach, P. M., & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument and Computation*, *4*(1), 64–88.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1), 3–44.

Geis, M. L., & Zwicky, A. M. (1971). On invited inferences. *Linguistic inquiry*, *2*(4), 561–566.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Grice, H. (1989). Indicative conditionals. In *Studies in the way of words* (pp. 58–85).

Grusdt, B., Lassiter, D., & Franke, M. (2021). *Probabilistic modeling of rational communication with conditionals.*

Herbstritt, M., & Franke, M. (2019, mar). Complex probability expressions & higher-order uncertainty: Compositional semantics, probabilistic pragmatics & experimental data. *Cognition*, *186*, 50–71.

Horn, L. R. (1989). *A natural history of negation*. Chicago: Chicago University Press.

Horn, L. R. (2000). From if to iff: Conditional perfection as pragmatic strengthening. *Journal of Pragmatics*, *32*(3), 289-326.

Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological review*, *109*, 646.

Krynski, T. R., & Tenenbaum, J. B. (2007). The Role of Causality in Judgment Under Uncertainty. *Journal of Experimental Psychology: General*, *136*(3).

Lewis, D. (1973). *Counterfactuals*. Harvard University Press.

Simons, M. (2001). Disjunction and alternativeness. *Linguistics and Philosophy*, *24*(5), 597–619.

Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (Vol. 2, pp. 98–112). Oxford University Press.

Veltman, F. (1986). Data semantics and the pragmatics of conditionals. In E. C. Traugott, A. ter Meulen, J. S. Reilly, & C. A. Ferguson (Eds.), *On conditionals* (pp. 147–168). Cambridge University Press.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

[7] In 157 trials conditionals were used, 151 from these in *dependent* situations, and in 39.7% of these, the probability for the antecedent-block to fall was assigned a larger probability than 0.7, and similarly in 40.4%, this probability was rated larger than the respective participant's average literal meaning threshold, indicating that they were, despite the utterance of a conditional, quite confident about the antecedent-block to fall.